

recombination into the chromosome.

Locked in prophage are useful for mapping of genes, operons, and/or specific mutations with either desirable or undesirable phenotypes. Locked-in prophages can also provide a means to separate and map multiple mutations in a given host. If one is looking for beneficial mutations outside a gene or operon of interest, then an unmodified gene or operon can be transduced into a mutagenized or stochastic &/or non-stochastic mutagenized host then screened for the presence of desired secondary mutations. Alternatively, the gene/operon of interest can be readily moved from a mutagenized/stochastic &/or non-stochastic mutagenized host into a different background to screen/select for modifications in the gene/operon itself. It is also possible to develop similar genetic elements using other combinations of transposable elements and bacteriophages or viruses as well. Similar systems are set up in other organisms, e.g., that do not allow replication of P22 or Pl. Broad host range phages and transposable elements are especially useful. Similar genetic elements are derived from other temperate phages that also package by a headful mechanism. In general, these are the phages that are capable of generalized transduction. Viruses infecting eukaryotic cells may be adapted for similar purposes. Examples of generalized transducing phages that are useful are described in: Green et al., "Isolation and preliminary characterization of lytic and lysogenic phages with wide host range within the streptomycetes", J Gen Microbiol 131(9):2459-2465 (1985); Studdard et al., "Genome structure in Streptomyces spp.: adjacent genes on the S. coelicolor A3(2) linkage map have cotransducible analogs in S. venezuelae", J Bacteriol 169(8):3 814-3 816 (1987); Wang et al., "High frequency generalized transduction by miniMu plasmid phage", Genetics 116(2):201-206, (1987); Welker, N. E., "Transduction in *Bacillus stearothermophilus*", J Bacteriol, 176(11):3354-3359, (1988); Darzins et al., "Mini-D3112 bacteriophage transposable elements for genetic analysis of *Pseudomonas aeruginosa*", J Bacteriol 171(7):3909-3916 (1989); Hugouvieux-Cotte-Pattat et al., "Expanded linkage map of *Erwinia chrysanthemi* strain 3937", Mol Microbiol 3(5):573-581, (1989); Ichige et al. "Establishment of gene transfer systems for and construction of the genetic map of a marine Vibrio

strain", J Bacteriol 171(4):1825-1834 (1989); Murainatsu et al., "Two generalized transducing phages in *Vibrio parahaemolyticus* and *Vibrio alginolyticus*", Microbiol Immunol (12):1073-1084 (1991); Regue et al., "A generalized transducing bacteriophage for *Serratia marcescens*", Res Microbiol 42(I):23 -27, (1991) - Kiesel et al , "Phage Acm I -mediated transduction in the facultatively methanol-utilizing *Aacetobacter* *methanolicus* MB 58/4", J Gen Virol 74(9):1741-1745 (1993); Blahova et al., "Transduction of imipenem resistance by the phage F- 116 from a nosocomial strain of *Pseudomonas aeruginosa* isolated in Slovakia", Acta Virol 38(5):247-250 (1994); Kidambi et al., "Evidence for phage- mediated gene transfer among *Pseudomonas aeruginosa* strains on the phylloplane", Appl Environ Microbiol 60:(2)496-500 (1994); Weiss et al., "Isolation and characterization of a generalized transducing phage for *Xanthomonas campestris* pv. *campestris*", J Bacteriol 176(11):3354- 3359 (1994); Matsumoto et al., "Clustering of the trp genes in *Burkholderia* (formerly *Pseudomonas*) *cepacia*", FEMS Microbiol Lett 134(2-3):265-271 (1995); Schicklmaier et al., "Frequency of generalized transducing phages in natural isolates of the *Salmonella typhimurium* complex", Appl Environ Microbiol 61(4): 61(4):1637-1640 (1995); Humphrey et al., "Purification and characterization of VSH- 1, a generalized transducing bacteriophage of *Serpulina hyodysenteriae*", J Bacteriol 179(2):323-329 (1997); Willi et al, "Transduction of antibiotic resistance markers among *Actinobacillus actinomycetemcomitans* strains by temperate bacteriophages Aa phi 23 ", Cell Mol Life Sci 53 (11-12):904- 910 (1997); Jensen et al., "Prevalence of broad-host-range lytic bacteriophages of *Sphaerophilus natans*, *Escherichia coli*, and *Pseudomonas aeruginosa*", Appl Environ Microbiol 64(2):575-580 (1998), and Nedelmann el al., "Generalized transduction for genetic linkage analysis and transfer of transposon insertions in different *Staphylococcus epidermidis* strains", Zentiviralaffil Bakteriol 287(1-2):85-92 (1998).

A Mud-PI/Tn-P1 system comparable to Mud-P22 is developed using phage P 1. Phage P 1 has an advantage of packaging much larger (about 110 kb) fragments per headful. Phage P 1 is currently used to create bacterial artificial chromosomes or BAC's. P 1 -based BAC vectors are designed along these principles so that cloned

DNA is packaged into phage particles, rather than the current system, which requires DNA preparation from single-copy episomes. This combines the advantages of both systems in having the genes cloned in a stable single-copy format, while allowing for amplification and specific packaging of cloned DNA upon induction of the prophage.

4.6.1.2.1.16 RANDOM PLACEMENT OF GENES OR IMPROVED GENES THROUGHOUT THE GENOME FOR OPTIMIZATION OF GENE CONTEXT

The placement and orientation of genes in a host chromosome (the "context" of the gene in a chromosome) or episome has large effects on gene expression and activity. Random integration of plasmid or other episomal sequences into a host chromosome by non-homologous recombination, followed by selection or screening for the desired phenotype, is a preferred way of identifying optimal chromosomal positions for expression of a target. This strategy is illustrated herein.

A variety of transposon mediated delivery systems can be employed to deliver genes of interest, either individual genes, genomic libraries, or a library of stochastic &/or non-stochastic mutagenized gene(s) randomly throughout the genome of a host. Thus, in one preferred embodiment, the improvement of a cellular function is achieved by cloning a gene of interest, for example a gene encoding a desired metabolic pathway, within a transposon delivery vehicle.

Such transposon vehicles are available for both Gram-negative and Gram-positive bacteria. De Lorenzo and Timis (1994) Methods in Enzymology 235:385404 describe the analysis and construction of stable phenotypes in gram- negative Bacteria with Tn5- and Tn 10-derived minitransposons. Kleckner et al. (1991) Methods in Enzymology 204, chapter 7 describe uses of transposons such as Tn 10, including for use in gram positive bacteria. Petit el al. (1990) Journal of Bacteriology, 172(12):6736-6740 describe Tn10 derived transposons active in *Bacillus Subtilis*. The

transposon delivery vehicle is introduced into a cell population, which is then selected for recombinant cells that have incorporated the transposon into the genome.

The selection is typically by any of a variety of drug resistant markers also carried within the transposon. The selected subpopulation is screened for cells having improved expression of the gene(s) of interest. Once cells harboring the genes of interest in the optimal location are isolated, the genes are amplified from within the genome using PCR, stochastic &/or non-stochastic mutagenized, and cloned back into a similar transposon delivery vehicle which contains a different selection marker within the transposon and lacks the transposon integrase gene.

This stochastic &/or non-stochastic mutagenized library is then transformed back into the strain harboring the original transposon, and the cells are selected for the presence of the new resistance marker and the loss of the previous selection marker. Selected cells are enriched for those that have exchanged by homologous recombination the original transposon for the new transposon carrying members of the stochastic &/or non-stochastic mutagenized library. The surviving cells are then screened for further improvements in the expression of the desired phenotype. The genes from the improved cells are then amplified by the PCR and stochastic &/or non-stochastic mutagenized again. This process is carried out recursively, oscillating each cycle between the different selection markers. Once the gene(s) of interest are optimized to a desired level, the fragment can be amplified and again randomly distributed throughout the genome as described above to identify the optimal location of the improved genes.

Alternatively, the gene(s) conferring a desired property may not be known. In this case the DNA fragments cloned within the transposon delivery vehicle could be a library of genomic fragments originating from a population of cells derived from one or more strains having the desired property(ies). The library is delivered to a

population of cells derived from one or more strains having or lacking the desired property(ies) and cells incorporating the transposon are selected. The surviving cells are then screened for acquisition or improvement of the desired property. The fragments contained within the surviving cells are amplified by PCR and then cloned as a pool into a similar transposon delivery vector harboring a different selection marker from the first delivery vector. This library is then delivered to the pool of surviving cells, and the population having acquired the new selective marker is selected. The selected cells are then screened for further acquisition or improvement of the desired property.

In this way the different possible combinations of genes conferring or improving a desired phenotype are explored in a combinatorial fashion. This process is carried out repetitively with each new cycle employing an additional selection marker. Alternatively, PCR fragments are cloned into a pool of transposon vectors having different selective markers. These are delivered to cells and selected for 1, 2, 3, or more markers.

Alternatively, the amplified fragments from each improved cell are stochastic &/or non-stochastic mutagenized independently. The stochastic &/or non-stochastic mutagenized libraries are then cloned back into a transposon delivery vehicle similar to the original vector but containing a different selection marker and lacking the transposase gene. Selection is then for acquisition of the new marker and loss of the previous marker. Selected cells are enriched for those incorporating the stochastic &/or non-stochastic mutagenized variants of the amplified genes by homologous recombination. This process is carried out recursively, oscillating each cycle between the two selective markers.

4.6.1.2.1.17 IMPROVEMENT OF OVEREXPRESSED GENES FOR A DESIRED PHENOTYPE

The improvement of a cellular property or phenotype is often enhanced by increasing the copy number or expression of gene(s) participating in the expression of that property. Genes that have such an effect on a desired property can also be improved by DNA stochastic &/or non-stochastic mutagenesis to have a similar effect. A genomic DNA library is cloned into an overexpression vector and transformed into a target cell population such that the genomic fragments are highly expressed in cells selected for the presence of the overexpression vector. The selected cells are then screened for improvement of a desired property. The overexpression vector from the improved cells are isolated and the cloned genomic fragments stochastic &/or non-stochastic mutagenized. The genomic fragment carried in the vector from each improved isolate is stochastic &/or non-stochastic mutagenized independently or with identified homologous genes (family stochastic &/or non-stochastic mutagenesis). The stochastic &/or non-stochastic mutagenized libraries are then delivered back to a population of cells and the selected transformants rescreened for further improvements in the desired property. This stochastic &/or non-stochastic mutagenesis/screening process is cycled recursively until the desired property has been optimized to the desired level. As stated above, gene dosage can greatly enhance a desired cellular property.

One method of increasing gene copy number of unknown genes is using a method of random amplification (see also, Mavingui et. al. (1997) Nature Biotech, 15, 5 64). In this method, a genomic library is cloned into a suicide vector containing a selective marker that also at higher dosage provides an enhanced phenotype. An example of such a marker is the kanamycin resistance gene. At successively higher copy number, resistance to successively higher levels of kanamycin is achieved. The genomic library is delivered to a target cell by any of a variety of methods including transformation, transduction, conjugation, etc. Cells that have incorporated the vector into the chromosome by homologous recombination between the vector and chromosomal copies of the cloned genes can be selected by requiring expression of the selection marker under conditions where the vector does not replicate. This recombination event results in the duplication of the cloned DNA fragment in the host

chromosome with a copy of the vector and selection marker separating the two copies. The population of surviving cells are screened for improvement of a desired cellular property resulting from the gene duplication event. Further gene duplication events resulting in additional copies of the original cloned DNA fragments can be generated by further propagating the cells under successively more stringent selective conditions i.e. increased concentrations of kanamycin. In this case selection requires increased copies of the selective marker, but increased copies of the desired gene fragment is also concomitant. Surviving cells are further screened for an improvement in the desired phenotype. The resulting population of cells likely resulted in the amplification of different genes since often many genes effect a given phenotype. To generate a library of the possible combinations of these genes, the original selected library showing phenotypic improvements are recombined, using the methods described herein, e.g., protoplast fusion, split pool transduction, transformation, conjugation, etc.

The recombined cells are selected for increased expression of the selective marker. Survivors are enriched for cells having incorporated additional copies of the vector sequence by homologous recombination, and these cells will be enriched for those having combined duplications of different genes. In other words, the duplication from one cell of enhanced phenotype becomes combined with the duplication of another cell of enhanced phenotype. These survivors are screened for further improvements in the desired phenotype. This procedure is repeated recursively until the desired level of phenotypic expression is achieved.

Alternatively, genes that have been identified or are suspected as being beneficial in increased copy number are cloned in tandem into appropriate plasmid vectors. These vectors are then transformed and propagated in an appropriate host organism. Plasmid-plasmid recombination between the cloned gene fragments result in further duplication of the genes. Resolution of the plasmid doublet can result in the uneven distribution of the gene copies, with some plasmids having additional gene copies and

others having fewer gene copies. Cells carrying this distribution of plasmids are then screened for an improvement in the phenotype effected by the gene duplications.

In summary, a method of selecting for increased copy number of a nucleic acid sequence by the above procedure is provided. In the method, a genomic library in a suicide vector comprising a dose-sensitive selectable marker is provided, as noted above. The genomic library is transduced into a population of target cells. The target cells are selected in a population of target cells for increasing doses of the selectable marker under conditions in which the suicide vector does not replicate episomally. A plurality of target cells are selected for the desired phenotype, recombined and reselected. The process is recursively repeated, if desired, until the desired phenotype is obtained.

4.6.1.2.1.18 STRATEGIES FOR IMPROVING GENOMIC STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS VIA TRANSFORMATION OF LINEAR DNA FRAGMENTS

Wild-type members of the Enterobacteriaceae (e.g., *Escherichia coli*) are typically resistant to genetic exchange following transformation of linear DNA molecules.

This is due, at least in part, to the Exonuclease V (Exo V) activity of the RecBCD holoenzyme which rapidly degrades linear DNA molecules following transformation. Production of ExoV has been traced to the *recD* gene, which encodes the D subunit of the holoenzyme. As demonstrated by Russel et al. (1989) Journal of Bacteriology 2609-2613, homologous recombination between a transformed linear donor DNA molecule and the chromosome of recipient is readily detected in a strains bearing a loss of function mutation in a *recD* mutant.

The use of *recD* strains provides a simple means for genomic stochastic &/or non-

stochastic mutagenesis of the Enterobacteriaceae. For example, a bacterial strain or set of related strains bearing a recD null mutation (e.g., the *E. coli* recD1903::mini-Tet allele) is mutagenized and screened for the desired properties. In a split-pool fashion, chromosomal DNA prepared on one aliquot could be used to transform (e.g., via electroporation or chemically induced competence) the second aliquot. The resulting transformants are then screened for improvement, or recursively transformed prior to screening.

The use of RecE/ recT as described supra, can improve homologous recombination of linear DNA fragments. The RecBCD holoenzyme plays an important role in initiation of RecA-dependent homologous recombination. Upon recognizing a dsDNA end, the RecBCD enzyme unwinds and degrades the DNA asymmetrically in a 5' to 3' direction until it encounters a chi (or 'X')-site (consensus 5'-GCTGGTGG-3) which attenuates the nuclease activity. This results in the generation of a ssDNA terminating near the c site with a 3'-ssDNA tail that is preferred for RecA loading and subsequent invasion of dsDNA for homologous recombination. Accordingly, preprocessing of transforming fragments with a 5' to 3' specific ssDNA Exonuclease, such as Lamda () exonuclease (available, e.g., from Boeringer Mannheim) prior to transformation may serve to stimulate homologous recombination in recD- strain by providing ssDNA invasive end for RecA loading and subsequent strand invasion.

The addition of DNA sequence encoding chi-sites (consensus 5'-GCTGGTGG-3') to DNA fragments can serve to both attenuate Exonuclease V activity and stimulate homologous recombination, thereby obviating the need for a recD mutation (see also, Kowalczykowski, et al. (1994) "Biochemistry of homologous recombination in *Escherichia coli*," *Microbiol. Rev.* 58:401-465 and Jessen, et al. (1998) "Modification of bacterial artificial chromosomes through Chi-stimulated homologous recombination and its application in zebrafish transgenesis." *Proc. Natl. Acad. Sci.* 95:5121- 5126). Chi sites are optionally included in linkers ligated to the ends of transforming fragments or incorporated into the external primers used to generate

DNA fragments to be transformed. The use of recombination-stimulatory sequences such as chi is a generally useful approach for evolution of a broad range of cell types by fragment transformation. Methods to inhibit or mutate analogs of Exo V or other nucleases (such as, Exonucleases I (endA 1), 111 (nth), IV (nfo), VII, and VIII of *E. coli*) is similarly useful.

Inhibition or elimination of nucleases, or modification of ends of transforming DNA fragments to render them resistant to exonuclease activity has applications in evolution of a broad range of cell types.

4.6.1.2.1.19 STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS TO OPTIMIZE UNKNOWN INTERACTIONS

Many observed traits are the result of complex interactions of multiple genes or gene products. Most such interactions are still uncharacterized. Accordingly, it is often unclear which genes need to be optimized to achieve a desired trait, even if some of the genes contributing to the trait are known.

This lack of characterization is not an issue during DNA stochastic &/or non-stochastic mutagenesis, which produces solutions that optimize whatever is selected for. An alternative approach, which has the potential to solve not only this problem, but also anticipated future rate limiting factors, is complementation by overexpression of unknown genomic sequences.

A library of genomic DNA is first made as described, supra. This is transformed into the cell to be optimized and transformants are screened for increases in a desired property. Genomic fragments which result in an improved property are evolved by DNA stochastic &/or non-stochastic mutagenesis to further increase their beneficial effect. This approach requires no sequence information, nor any knowledge or

assumptions about the nature of protein or pathway interactions, or even of what steps are rate -limiting; it relies only on detection of the desired phenotype. This sort of random cloning and subsequent evolution by DNA stochastic &/or non-stochastic mutagenesis of positively interacting genomic sequences is extremely powerful and generic. A variety of sources of genomic DNA are used, from isogenic strains to more distantly related species with potentially desirable properties. In addition, the technique is applicable to any cell for which the molecular biology basics of transformation and cloning vectors are available, and for any property which can be assayed (preferably in a high-throughput format). Alternatively, once optimized, the evolved DNA can be returned to the chromosome by homologous recombination or randomly by phage mediated site-specific recombination.

4.6.1.2.1.20 HOMOLOGOUS RECOMBINATION WITHIN THE CHROMOSOME

Homologous recombination within the chromosome is used to circumvent the limitations of plasmid based evolution and size restrictions. The strategy is similar to that described above for stochastic &/or non-stochastic mutagenesis genes within their chromosomal context, except that no *in vitro* stochastic &/or non-stochastic mutagenesis occurs. Instead, the parent strain is treated with mutagens such as ultraviolet light or nitrosoguanidine, and improved mutants are selected. The improved mutants are pooled and split. Half of the pool is used to generate random genomic fragments for cloning into a homologous recombination vector. Additional genomic fragments are optionally derived from related species with desirable properties. The cloned genomic fragments are homologously recombined into the genomes of the remaining half of the mutant pool, and variants with improved properties are selected. These are subjected to a further round of mutagenesis, selection and recombination. Again this process is entirely generic for the improvement of any whole cell biocatalyst for which a recombination vector and an assay can be developed. Here again, it should be noted that recombination can be performed recursively prior to screening.

4.6.1.2.1.21 METHODS FOR RECURSIVE SEQUENCE RECOMBINATION

As shown herein, DNA Stochastic &/or non-stochastic mutagenesis provides most rapid technology for evolution of complex new functions. As shown herein, recombination in DNA stochastic &/or non-stochastic mutagenesis achieves accumulation of multiple beneficial mutations in a few cycles. In contrast, because of the high frequency of deleterious mutations relative to beneficial ones, iterative point mutation must build beneficial mutations one at a time, and consequently requires many cycles to reach the same point. As shown herein, rather than a simple linear sequence of mutation accumulation, DNA stochastic &/or non-stochastic mutagenesis is a parallel process where multiple problems may be solved independently, and then combined.

4.6.1.2.2 IN VIVO FORMATS

4.6.1.2.2.1 PLASMID-PLASMID RECOMBINATION

The initial substrates for recombination are a collection of polynucleotides comprising variant forms of a gene. The variant forms usually show substantial sequence identity to each other sufficient to allow homologous recombination between substrates. The diversity between the polynucleotides can be natural (e.g., allelic or species variants), induced (e.g., error-prone PCR or error-prone recursive sequence recombination), or the result of in vitro recombination. Diversity can also result from resynthesizing genes encoding natural proteins with alternative codon usage. There should be at least sufficient diversity between substrates that recombination can generate more diverse products than there are starting materials. There must be at least two substrates differing in at least two positions.

However, commonly a library of substrates of 10^3 - 10^8 members is employed. The degree of diversity depends on the length of the substrate being recombined and the extent of the functional change to be evolved. Diversity at between 0.1-25% of

positions is typical. The diverse substrates are incorporated into plasmids. The plasmids are often standard cloning vectors, e.g., bacterial multicopy plasmids. However, in some methods to be described below, the plasmids include mobilization (MOB) functions. The substrates can be incorporated into the same or different plasmids. Often at least two different types of plasmid having different types of selectable markers are used to allow selection for cells containing at least two types of vector. Also, where different types of plasmid are employed, the different plasmids can come from two distinct incompatibility groups to allow stable co-existence of two different plasmids within the cell. Nevertheless, plasmids from the same incompatibility group can still co-exist within the same cell for sufficient time to allow homologous recombination to occur.

Plasmids containing diverse substrates are initially introduced into cells by any method (e.g., chemical transformation, natural competence, electroporation, biolistics, packaging into phage or viral systems). Often, the plasmids are present at or near saturating concentration (with respect to maximum transfection capacity) to increase the probability of more than one plasmid entering the same cell. The plasmids containing the various substrates can be transfected simultaneously or in multiple rounds. For example, in the latter approach cells can be transfected with a first aliquot of plasmid, transfectants selected and propagated, and then infected with a second aliquot of plasmid.

Having introduced the plasmids into cells, recombination between substrates to generate recombinant genes occurs within cells containing multiple different plasmids merely by propagating the cells. However, cells that receive only one plasmid are unable to participate in recombination and the potential contribution of substrates on such plasmids to evolution is not fully exploited (although these plasmids may contribute to some extent if they are propagated in mutator cells). The rate of evolution can be increased by allowing all substrates to participate in recombination. Such can be achieved by subjecting transfected cells to electroporation. The

conditions for electroporation are the same as those conventionally used for introducing exogenous DNA into cells (e.g., 1,000-2,500 volts, 400 uF and a 1-2 mM gap). Under these conditions, plasmids are exchanged between cells allowing all substrates to participate in recombination.

In addition the products of recombination can undergo further rounds of recombination with each other or with the original substrate. The rate of evolution can also be increased by use of conjugative transfer. To exploit conjugative transfer, substrates can be cloned into plasmids having MOB genes, and tra genes are also provided in cis or in trans to the MOB genes. The effect of conjugative transfer is very similar to electroporation in that it allows plasmids to move between cells and allows recombination between any substrate and the products of previous recombination to occur, merely by propagating the culture. The rate of evolution can also be increased by fusing cells to induce exchange of plasmids or chromosomes. Fusion can be induced by chemical agents, such as PEG, or viral proteins, such as influenza virus hemagglutinin, HSV-1 gB and gD. The rate of evolution can also be increased by use of mutator host cells (e.g., Mut L, S, D, T, H in bacteria and Ataxia telangiectasia human cell lines) .

The time for which cells are propagated and recombination is allowed to occur, of course, varies with the cell type but is generally not critical, because even a small degree of recombination can substantially increase diversity relative to the starting materials. Cells bearing plasmids containing recombinant genes are subject to screening or selection for a desired function. For example, if the substrate being evolved contains a drug resistance gene, one would select for drug resistance. Cells surviving screening or selection can be subjected to one or more rounds of screening/selection followed by recombination or can be subjected directly to an additional round of recombination. The next round of recombination can be achieved by several different formats independently of the previous round. For example, a further round of recombination can be effected simply by resuming the

electroporation or conjugation-mediated intercellular transfer of plasmids described above.

Alternatively, a fresh substrate or substrates, the same or different from previous substrates, can be transfected into cells surviving selection/screening. Optionally, the new substrates are included in plasmid vectors bearing a different selective marker and/or from a different incompatibility group than the original plasmids. As a further alternative, cells surviving selection/screening can be subdivided into two subpopulations, and plasmid DNA from one subpopulation transfected into the other, where the substrates from the plasmids from the two subpopulations undergo a further round of recombination. In either of the latter two options, the rate of evolution can be increased by employing DNA extraction, electroporation, conjugation or mutator cells, as described above. In a still further variation, DNA from cells surviving screening/selection can be extracted and subjected to *in vitro* recursive sequence recombination. After the second round of recombination, a second round of screening/selection is performed, preferably under conditions of increased stringency. If desired, further rounds of recombination and selection/screening can be performed using the same strategy as for the second round.

With successive rounds of recombination and selection/screening, the surviving recombined substrates evolve toward acquisition of a desired phenotype. Typically, in this and other methods of recursive recombination, the final product of recombination that has acquired the desired phenotype differs from starting substrates at 0.1%-50% of positions and has evolved at a rate orders of magnitude in excess (e.g., by at least 10-fold, 100-fold, 1000-fold, or 10,000 fold) of the rate of naturally acquired mutation of about 1 mutation per 10^9 positions per generation (see Anderson et al. Proc. Natl. Acad. Sci. U.S.A. 93:906-907 (1996)). The "final product" may be transferred to another host more desirable for utilization of the "stochastic &/or non-stochastic mutagenized" DNA.

This is particularly advantageous in situations where the more desirable host is less efficient as a host for the many cycles of mutation/recombination due to the lack of molecular biology or genetic tools available for other organisms such as *E. coli*.

4.6.1.2.2.2 VIRUS-PLASMID RECOMBINATION

The strategy used for plasmid-plasmid recombination can also be used for virus-plasmid recombination; usually, phage-plasmid recombination. However, some additional comments particular to the use of viruses are appropriate.

The initial substrates for recombination are cloned into both plasmid and viral vectors. It is usually not critical which substrate(s) are inserted into the viral vector and which into the plasmid, although usually the viral vector should contain different substrate(s) from the plasmid. As before, the plasmid (and the virus) typically contains a selective marker.

The plasmid and viral vectors can both be introduced into cells by transfection as described above. However, a more efficient procedure is to transfet the cells with plasmid, select transfectants and infect the transfectants with virus. Because the efficiency of infection of many viruses approaches 100% of cells, most cells transfected and infected by this route contain both a plasmid and virus bearing different substrates.

Homologous recombination occurs between plasmid and virus generating both recombined plasmids and recombined virus. For some viruses, such as filamentous phage, in which intracellular DNA exists in both double-stranded and single-stranded forms, both can participate in recombination.

Provided that the virus is not one that rapidly kills cells, recombination can be augmented by use of electroporation or conjugation to transfer plasmids between cells. Recombination can also be augmented for some types of virus by allowing the progeny virus from one cell to reinfect other cells. For some types of virus, virus infected-cells show resistance to superinfection. However, such resistance can be overcome by infecting at high multiplicity and/or using mutant strains of the virus in which resistance to superinfection is reduced.

The result of infecting plasmid-containing cells with virus depends on the nature of the virus. Some viruses, such as filamentous phage, stably exist with a plasmid in the cell and also extrude progeny phage from the cell. Other viruses, such as lambda having a cosmid genome, stably exist in a cell like plasmids without producing progeny virions.

Other viruses, such as the T-phage and lytic lambda, undergo recombination with the plasmid but ultimately kill the host S cell and destroy plasmid DNA. For viruses that infect cells without killing the host, cells containing recombinant plasmids and virus can be screened/selected using the same approach as for plasmid-plasmid recombination. Progeny virus extruded by cells surviving selection/screening can also be collected and used as substrates in subsequent rounds of recombination. For viruses that kill their host cells, recombinant genes resulting from recombination reside only in the progeny virus. If the screening or selective assay requires expression of recombinant genes in a cell, the IS recombinant genes should be transferred from the progeny virus to another vector, e.g., a plasmid vector, and retransfected into cells before selection/screening is performed.

For filamentous phage, the products of recombination are present in both cells surviving recombination and in phage extruded from these cells. The dual source of recombinant products provides some additional options relative to the plasmid-plasmid recombination. For example, DNA can be isolated from phage particles for use in a round of in vitro recombination. Alternatively, the progeny 2S phage can be used to transfect or infect cells surviving a previous round of screening/selection, or fresh cells transfected with fresh substrates for recombination.

4.6.1.2.2.3 VIRUS-VIRUS RECOMBINATION

The principles described for plasmid-plasmid and plasmid-viral recombination can be applied to virus-virus recombination with a few modifications. The initial substrates for recombination are cloned into a viral vector. Usually, the same vector is used for

all substrates.

Preferably, the virus is one that, naturally or as a result of mutation, does not kill cells. After insertion, some viral genomes can be packaged in vitro or using a packaging cell line. The packaged viruses are used to infect cells at high multiplicity such that there is a high probability that a cell will receive multiple viruses bearing different substrates.

After the initial round of infection, subsequent steps depend on the nature of infection as discussed in the previous section. For example, if the viruses have phagemid genomes such as lambda cosmids or M13, F1 or Fd phagemids, the phagemids behave as plasmids within the cell and undergo recombination simply by propagating the cells. Recombination is particularly efficient between single-stranded forms of intracellular DNA. Recombination can be augmented by electroporation of cells.

Following selection/screening, cosmids containing recombinant genes can be recovered from surviving cells, e.g., by heat induction of a cos- lysogenic host cell, or extraction of DNA by standard procedures, followed by repackaging cosmid DNA in vitro.

If the viruses are filamentous phage, recombination of replicating form DNA occurs by propagating the culture of infected cells. Selection/screening identifies colonies of cells containing viral vectors having recombinant genes with improved properties, together with phage extruded from such cells. Subsequent options are essentially the same as for plasmid-viral recombination.

4.6.1.2.2.4 CHROMOSOME RECOMBINATION

This format can be used to especially evolve chromosomal substrates. The format is particularly useful in situations in which many chromosomal genes contribute to a phenotype or one does not know the exact location of the chromosomal gene(s) to be

evolved. The initial substrates for recombination are cloned into a plasmid vector. If the chromosomal gene(s) to be evolved are known, the substrates constitute a family of sequences showing a high degree of sequence identity but some divergence from the chromosomal gene. If the chromosomal genes to be evolved have not been located, the initial substrates usually constitute a library of DNA segments of which only a small number show sequence identity to the gene or gene(s) to be evolved. Divergence between plasmid-borne substrate and the chromosomal gene(s) can be induced by mutagenesis or by obtaining the plasmid- borne substrates from a different species than that of the cells bearing the chromosome.

The plasmids bearing substrates for recombination are transfected into cells having chromosomal gene(s) to be evolved. Evolution can occur simply by propagating the culture, and can be accelerated by transferring plasmids between cells by conjugation or electroporation. Evolution can be further accelerated by use of mutator host cells or by seeding a culture of nonmutator host cells being evolved with mutator host cells and inducing intercellular transfer of plasmids by electroporation or conjugation. Preferably, mutator host cells used for seeding contain a negative selectable marker to facilitate isolation of a pure culture of the nonmutator cells being evolved. Selection/screening identifies cells bearing chromosomes and/or plasmids that have evolved toward acquisition of a desired function.

Subsequent rounds of recombination and selection/screening proceed in similar fashion to those described for plasmid-plasmid recombination. For example, further recombination can be effected by propagating cells surviving recombination in combination with electroporation or conjugative transfer of plasmids. Alternatively, plasmids bearing additional substrates for recombination can be introduced into the surviving cells. Preferably, such plasmids are from a different incompatibility group and bear a different selective marker than the original plasmids to allow selection for cells containing at least two different plasmids. As a further alternative, plasmid and/or chromosomal DNA can be isolated from a subpopulation of surviving cells and transfected into a second subpopulation. Chromosomal DNA can be cloned into a plasmid vector before transfection.

4.6.1.2.2.5 VIRUS-CHROMOSOME RECOMBINATION

As in the other methods described above, the virus is usually one that does not kill the cells, and is often a phage or phagemid. The procedure is substantially the same as for plasmid-chromosome recombination. Substrates for recombination are cloned into the vector. Vectors including the substrates can then be transfected into cells or in vitro packaged and introduced into cells by infection. Viral genomes recombine with host chromosomes merely by propagating a culture. Evolution can be accelerated by allowing intercellular transfer of viral genomes by electroporation, or reinfection of cells by progeny virions. Screening/selection identifies cells having chromosomes and/or viral genomes that have evolved toward acquisition of a desired function.

There are several options for subsequent rounds of recombination. For example, viral genomes can be transferred between cells surviving selection/recombination by electroporation. Alternatively, viruses extruded from cells surviving selection/screening can be pooled and used to superinfect the cells at high multiplicity. Alternatively, fresh substrates for recombination can be introduced into the cells, either on plasmid or viral vectors.

4.6.1.2.2.6 POOLWISE WHOLE GENOME RECOMBINATION

Asexual evolution is a slow and inefficient process. Populations move as individuals rather than as a group. A diverse population is generated by mutagenesis of a single parent, resulting in a distribution of fit and unfit individuals. In the absence of a sexual cycle, each piece of genetic information for the surviving population remains in the individual mutants. Selection of the fittest results in many fit individuals being discarded, along with the genetically useful information they carry. Asexual evolution proceeds one genetic event at a time, and is thus limited by the intrinsic value of a single genetic event. Sexual evolution moves more quickly and efficiently. Mating within a population consolidates genetic information within the population and results in useful information being combined together.

The combining of useful genetic information results in progeny that are much more fit than their parents. Sexual evolution thus proceeds much faster by multiple genetic events. These differences are further illustrated herein. In contrast to sexual evolution, DNA stochastic &/or non-stochastic mutagenesis is the recursive mutagenesis, recombination, and selection of DNA sequences.

Sexual recombination in nature effects pairwise recombination and results in progeny that are genetic hybrids of two parents. In contrast, DNA stochastic &/or non-stochastic mutagenesis in vitro effects poolwise recombination, in which progeny are hybrids of multiple parental molecules. This is because DNA stochastic &/or non-stochastic mutagenesis effects many individual pairwise recombination events with each thermal cycle. After many cycles the result is a repetitively inbred population, with the "progeny" being the F_X (for X cycles of stochastic &/or non-stochastic mutagenesis) of the original parental molecules. These progeny are potentially descendants of many or all of the original parents. One can graph to show a plot of the potential number of mutations an individual can accumulate by sequential, pairwise and poolwise recombination.

Poolwise recombination is an important feature to DNA stochastic &/or non-stochastic mutagenesis in that it provides a means of generating a greater proportion of the possible combinations of mutations from a single "breeding" experiment. In this way, the "genetic potential" of a population can be readily assessed by screening the progeny of a single DNA shuffling experiment.

For example, if a population consists of 10 single mutant parents, there are $2^{10}=1024$ possible combinations of those mutations ranging from progeny having 0- 10 mutations. Of these 1024, only 56 will result from a single pairwise cross (i.e those having 0, 1, and 2 mutations). In nature the multiparent combinations will eventually arise after multiple random sexual matings, assuming no selection is imparted to

remove some mutations from the population. In this way, sex effects the consolidation and sampling of all useful mutant combinations possible within a population. For the purposes of directed evolution, having the greatest number of mutant combinations entering a screen or selection is desirable so that the best progeny (i.e., according to the selection criteria used in the selection screen) is identified in the shortest possible time.

One challenge to *in vivo* and whole genome stochastic &/or non-stochastic mutagenesis is devising methods for effecting poolwise recombination or multiple repetitive pairwise recombination events. In crosses with a single pairwise cross per cycle before screening, the ability to screen the "genetic potential" of the starting population is limited. For this reason, the rate of *in vivo* and whole genome stochastic &/or non-stochastic mutagenesis mediated cellular evolution would be facilitated by effecting poolwise recombination. Two strategies for poolwise recombination are described below (protoplast fusion and transduction).

4.6.1.2.2.7 PROTOPLAST FUSION

Protoplast fusion (discussed *supra*) mediated whole genome stochastic &/or non-stochastic mutagenesis is one format that can directly effect poolwise recombination. Whole genome stochastic &/or non-stochastic mutagenesis is the recursive recombination of whole genomes, in the form of one or more nucleic acid molecule(s) (fragments, chromosomes, episomes, etc), from a population of organisms, resulting in the production of new organisms having distributed genetic information from at least two of the starting population of organisms. The process of protoplast fusion is further illustrated in herein.

Progeny resulting from the fusion of multiple parent protoplasts have been observed (Hopwood & Wright, 1978), however, these progeny are rare (10^{-4} - 10^{-6}). The low frequency is attributed to the distribution of fusants arising from two, three, four, etc

parents and the likelihood of the multiple recombination events (6 crossovers for a four parent cross) that would have to occur for multiparent progeny to arise. Thus, it is useful to enrich for the multiparent progeny. This can be accomplished, e.g., by repetitive fusion or enrichment for multiply fused protoplasts. The process of poolwise fusion and recombination is further illustrated herein.

4.6.1.2.8 REPETITIVE FUSION

Protoplasts of identified parental cells are prepared, fused and regenerated. Protoplasts of the regenerated progeny are then, without screening or enrichment, formed, fused and regenerated. This can be carried out for two, three, or more cycles before screening to increase the representation of multiparent progeny. The number of possible mutations/progeny doubles for each cycle. For example, if one cross produces predominantly progeny with 0, 1, and 2 mutations, a breeding of this population with itself will produce progeny with 0, 1, 2, 3, and 4 mutations, the third cross up to eight, etc. The representation of the multiparent progeny from these subsequent crosses will not be as high as the single and double parent progeny, but it will be detectable and much higher than from a single cross. The repetitive fusion prior to screening is analogous to many sexual crosses within a population, and the individual thermal cycles of in vitro DNA stochastic &/or non-stochastic mutagenesis described supra. A factor effecting the value of this approach is the starting size of the parental population. As the population grows, it becomes more likely that a multiparent fusion will arise from repetitive fusions. For example, if 4 parents are fused twice, the 4 parent progeny will make up approximately 0.2% of the total progeny. This is sufficient to find in a population of 3000 (95% confidence), but better representation is preferable. If ten parents are fused twice >20% of the progeny will be four parent offspring.

4.6.1.2.9 ENRICHMENT FOR MULTIPLE FUSED PROTOPLASTS

After the fusion of a population of protoplasts, the fusants are typically diluted into

hypotonic medium, to dilute out the fusing agent (e.g., 50% PEG). The fused cells can be grown for a short period to regenerate cell walls or separated directly and are then separated on the basis of size. This is carried out, e.g., by cell sorting, using light dispersion as an estimate of size, to isolate the largest fusants. Alternatively the fusants can be sorted by FACS on the basis of DNA content. The large fusants or those containing more DNA result from the fusion of multiple parents and are more likely to segregate to multiparent progeny. The enriched fusants are regenerated and screened directly or the progeny are fused recursively as above to further enrich the population for diverse mutant combinations.

4.6.1.2.2.10 TRANSDUCTION

Transduction can theoretically effect poolwise recombination, if the transducing phage particles contain predominantly host genomic DNA rather than phage DNA. If phage DNA is overly represented, then most cells will receive at least one undesired phage genome.

Phage particles generated from locked-in-prophage (supra) are useful for this purpose. A population of cells is infected with an appropriate transducing phage, and the lysate is collected and used to infect the same starting population. A high multiplicity of infection is employed to deliver multiple genomic fragments to each infected cell, thereby increasing the chance of producing recombinants containing mutations from more than two parent genomes.

The resulting transductants are recovered under conditions where phage can not propagate e.g., in the presence of citrate. This population is then screened directly or infected again with phage, with the resulting transducing particles being used to transduce the first progeny. This would mimic recursive protoplast fusion, multiple sexual recombination, and in vitro DNA stochastic &/or non-stochastic mutagenesis.

4.6.1.2.2.11 METHODS FOR WHOLE GENOME STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS BY BLIND FAMILY STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS OF PARSED GENOMES AND RECURSIVE CYCLES OF FORCED INTEGRATION AND EXCISION BY HOMOLOGOUS RECOMBINATION, AND SCREENING FOR IMPROVED PHENOTYPES

In vitro methods have been developed to reassemble single genes and operons, as set forth, e.g., herein. "Family" stochastic &/or non-stochastic mutagenesis of homologous genes within species and from different species is also an effective methods for accelerating molecular evolution. This section describes additional methods for extending these methods such that they can be applied to whole genomes.

In some cases, the genes that encode rate limiting steps in a biochemical process, or that contribute to a phenotype of interest are known. This method can be used to target family stochastic &/or non-stochastic mutagenized libraries to such loci, generating libraries of organisms with high quality family stochastic &/or non-stochastic mutagenized libraries of alleles at the locus of interest. An example of such a gene would be the evolution of a host chaperonin to more efficiently chaperone the folding of an overexpressed protein in *E. coli*.

The goals of this process are to reassemble homologous genes from two or more species and to then integrate the stochastic &/or non-stochastic mutagenized genes into the chromosome of a target organism.

Integration of multiple stochastic &/or non-stochastic mutagenized genes at multiple loci can be achieved using recursive cycles of integration (generating duplications), excision (leaving the improved allele in the chromosome) and transfer of additional evolved genes by serially applying the same procedure.

In the first step, genes to be stochastic &/or non-stochastic mutagenized into suitable bacterial vectors are subcloned. These vectors can be plasmids, cosmids, BACS or the like. Thus, fragments from 100 bp to 100 kb can be handled. Homologous fragments are then "family stochastic &/or non-stochastic mutagenized" together (i.e. homologous fragments from different species or chromosomal locations are homologously recombined). As a simple case, homologs from two species (say, *E. coli* and *Salmonella*) are cloned, family stochastic &/or non-stochastic mutagenized in vitro and cloned into an allele replacement vector (e.g., a vector with a positively selectable marker, a negatively selectable marker and conditionally active origin of replication). The basic strategy for whole genome family stochastic &/or non-stochastic mutagenesis of parsed (subcloned) genomes is additionally set forth herein.

The vectors are transfected into *E. coli* and selected, e.g., for drug resistance. Most drug resistant cells should arise by homologous recombination between a family stochastic &/or non-stochastic mutagenized insert and a chromosomal copy of the cloned insert. Colonies with improved phenotype are screened (e.g., by mass spectroscopy for enzyme activity or small molecule production, or a chromogenic screen, or the like, depending on the phenotype to be assayed). Negative selection (i.e. sue selection) is imposed to force excision of tandem duplication. Roughly half C, of the colonies should retain the improved phenotype. Importantly, this process regenerates a "clean" chromosome in which the wild type locus is replaced with a family stochastic &/or non-stochastic mutagenized fragment that encodes a beneficial allele. Since the chromosome is "clean" (i.e., has no vector sequences), other improved alleles can also be moved into this point on the chromosome by homologous recombination.

Selection or screening for improved phenotype can occur either after step 3 or step 4. If selection or screening takes place after step 3, then the improved allele can be conveniently moved to other strains by, for example, P I transduction. One can then regenerate a strain containing the improved allele but lacking vector sequences by

"negative selection" against the suc marker. In subsequent rounds, independently identified improved variants of the gene can be sequentially moved into the improved strain (e.g., by P I transduction of the drug marked tandem duplication above). Transductants are screened for further improvement in phenotype by virtue of receiving the transduced tandem duplication, which itself contains the family stochastic &/or non-stochastic mutagenized genetic material. Negative selection is again imposed and the process of stochastic &/or non-stochastic mutagenesis the improved strain is recursively repeated as desired.

Although this process was described with reference to targeting a gene or genes of interest, it can be used "blindly," making no assumptions about which locus is to be targeted. This procedure is set forth herein. For example, the whole genome of an organism of interest is cloned into manageable fragments (e.g., 10 kb for plasmid-based methods). Homologous fragments are then isolated from related species. Forced recombination with chromosomal homologs creates chimeras.

4.6.1.2.2.12 METHODS FOR HIGH THROUGHPUT FAMILY STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS OF GENES

For *E. coli*., cloning the genome in 10 kb fragments requires about 300 clones. The homologous fragments are isolated, e.g., from *Salmonella*. This gives roughly three hundred pairs of homologous fragments. Each pair is family stochastic &/or non-stochastic mutagenized and the stochastic &/or non-stochastic mutagenized fragments are cloned into an allele replacement vector. The inserts are integrated into the *E. coli* genome as described above. A global screen is made to identify variants with an improved phenotype. This serves as the basis collection of improvements that are to be stochastic &/or non-stochastic mutagenized to produce a desired strain. The stochastic &/or non-stochastic mutagenesis of these independently identified variants into one super strain is done as described above.

Family stochastic &/or non-stochastic mutagenesis has been shown to be an efficient method for creating high quality libraries of genetic variants. Given a cloned gene from one species, it is of interest to quickly and rapidly isolate homologs from other species, and this process can be rate limiting. For example, if one wants to perform family stochastic &/or non-stochastic mutagenesis on an entire genome, one may need to construct hundreds to thousands of individual family stochastic &/or non-stochastic mutagenized libraries.

In this embodiment, a gene of interest is optionally cloned into a vector in which ssDNA can be made. An example of such a vector is a phagemid vector with an M13 origin of replication. Genomic DNA or cDNA from a species of interest is isolated, denatured, annealed to the phagemid, and then enzymatically manipulated to clone it. The cloned DNA is then used to family reassemble with the original gene of interest. PCR based formats are also available. These formats require no intermediate cloning steps, and are, therefore, of particular interest for high throughput applications.

Alternatively, the gene of interest can be fished out using purified RecA protein. The gene of interest is PCR amplified using primers that are tagged with an affinity tag such as biotin, denatured, then coated with RecA protein (or an improved variant thereof). The coated ssDNA is then mixed with a gDNA plasmid library. Under the appropriate conditions, such as in the presence of non-hydrolyzable rATP analogs, RecA will catalyze the hybridization of the RecA coated gene (ssDNA) in the plasmid library. The heteroduplex is then affinity purified from the non-hybridizing plasmids of the gene library by adsorption of the labeled PCR products and its associated homologous DNA to an appropriate affinity matrix.

The homologous DNA is used in a family stochastic &/or non-stochastic mutagenesis reaction for improvement of the desired function. Stochastic &/or non-stochastic mutagenesis the E. coli chaperonin gene DnaJ with other homologs is described

below as an example. The example can be generalized to any other gene, including eukaryotic genes such as plant or animal genes (including mammalian genes), by following the format described.

As a first step, the *E. coli* DnaJ gene is cloned into an M13 phagemid vector. ssDNA is then produced, preferably in a dut(-) ung(-) strain so that Kunkel site directed mutagenesis protocols can be applied. Genomic DNA is then isolated from a non- *E. coli* source, such as *Salmonella* and *Yersinia Pestis*. The bacterial genomic DNAs are denatured and reannealed to the phagemid ssDNA (e.g., about 1 microgram of ssDNA). The reannealed product is treated with an enzyme such as Mung Bean nuclease that degrades ssDNA as an exonuclease but not as an endonuclease (the nuclease does not degrade mismatched DNA that is embedded in a larger annealed fragment). The standard Kunkel site directed mutagenesis protocol is used to extend the fragment and the target cells are transformed with the resulting mutagenized DNA.

In a first variation on the above, the procedure is adapted to the situation where the target gene or genes of interest are unknown. In this variation, the whole genome of the organism of interest is cloned in fragments (e.g., of about 10 kb each) into a phagemid. Single stranded phagemid DNA is then produced. Genomic DNA from the related species is denatured and annealed to the phagemids. Mung bean nuclease is used to trim away unhybridized DNA ends. Polymerase plus ligase is used to fill in the resulting gapped circles.

These clones are transformed into a mismatch repair deficient strain. When the mismatched molecules are replicated in the bacteria, most colonies contain both the *E. coli* and the homologous fragment. The two homologous genes are then isolated from the colonies (e.g., either by standard plasmid purification or colony PCR) and stochastic &/or non-stochastic mutagenized.

Another approach to generating chimeras that requires no in vitro stochastic &/or non-stochastic mutagenesis is simply to clone the *Salmonella* genome into an allele replacement vector, transform *E. coli*, and select for chromosomal integrants. Homologous recombination between *Salmonella* genes and *E. coli* homologs generate stochastic &/or non-stochastic mutagenized chimeras. A global screen is done to screen for improved phenotypes. Alternately, recursive transformation and recombination is performed to increase diversity prior to screening. If colonies with improved phenotypes are obtained, it is verified that the improvement is due to allele replacement by P I transduction into a fresh strain and counterscreening for improved phenotype. A collection of such improved alleles can then be combined into one strain using the methods for whole genome stochastic &/or non-stochastic mutagenesis by blind family stochastic &/or non-stochastic mutagenesis of parsed genomes as set forth herein. Additionally, once these loci are identified, it is likely that further rounds of stochastic &/or non-stochastic mutagenesis and screening will yield further improvements. This could be done by cloning the chimeric gene and then using the methods described in this disclosure to breed the gene with homologs from many different strains of bacteria.

In general, the transformants contain clones of the homologue of the target gene (e.g., *E. coli* DnaJ in the example above). Mismatch repair *in vivo* results in a decrease in diversity of the gene. There are at least two solutions to this. First, transduction can be performed into a mismatch repair deficient strain. Alternatively or in addition, the M13 template DNA can be selectively degraded, leaving the cloned homologue. This can be done using methods similar to the standard Eckstein site directed mutagenesis technique (General texts which describe general molecular biological techniques useful herein, including mutagenesis, include Sambrook et al., *Molecular Cloning - A Laboratory Manual* (2nd Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1989 ("Sambrook") and *Current Protocols in Molecular Biology*, F.M. Ausubel et al., eds., *Current Protocols*, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (supplemented through

1998) ("Ausubel")).

This method relies on incorporation of alpha thiol modified dNTPs during synthesis of the new strand followed by selective degradation of the template and resynthesis of the template strand. In one embodiment, the template strand is grown in a dut(-) ung(-) strain so that uracil is incorporated into the phagemid DNA. After extension as noted above (and before transformation) the DNA is treated with uracil glycosylate and an apurinic site endonuclease such as Endo III or Endo IV. The treated DNA is then treated with a processive exonuclease that resects from the resulting gaps while leaving the other strand intact (as in Eckstein mutagenesis). The DNA is polymerized and ligated. Target cells are then transformed. This process enriches for clones encoding the homologue which is not derived from the target (i.e., in the example above, the non- E. coli. homologue).

An analogous procedure is optionally performed in a PCR format. As applied to the DnaJ illustration above, DnaJ DNA is amplified by PCR with primers that build 30-mer priming sites on each end. The PCR is denatured and annealed with an excess of Salmonella genomic DNA. The Salmonella DnaJ gene hybridizes with the E. coli homologue. After treatment with Mung Bean nuclease, the resulting mismatched hybrid is PCR amplified with the flanking 30-mer primers. This PCR product can be used directly for family stochastic &/or non-stochastic mutagenesis. As genomics provides an increasing amount of sequence information, it is increasingly possible to directly PCR amplify homologs with designed primers. For example, given the sequence of the E. coli genome and of a related genome (i.e. Salmonella), each genome can be PCR amplified with designed primers in, e.g., 5 kb fragments. The homologous fragments can be put together in a pairwise fashion for stochastic &/or non-stochastic mutagenesis. For genome stochastic &/or non-stochastic mutagenesis, the stochastic &/or non-stochastic mutagenized products are cloned into the allele replacement vector and bred into the genome as described supra.

4.6.1.2.2.13 HYPER-RECOMBINOGENIC RecA CLONES

The invention further provides hyper-recombinogenic RecA proteins (see, the examples below). It is fully expected that one of skill can make a variety of related recombinogenic proteins given the disclosed sequences.

Standard molecular biological techniques can be used to make nucleic acids which comprise the given nucleic acids, e.g., by cloning the nucleic acids into any known vector. Examples of appropriate cloning and sequencing techniques, and instructions sufficient to direct persons of skill through many cloning exercises are found in Berger and Kimmel, Guide to Molecular Cloning Techniques, Methods in Enzymology volume 152 Academic Press, Inc., San Diego, CA (Berger); Sambrook et al. (1989) Molecular Cloning - A Laboratory Manual (2nd ed.) Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor Press, NY, (Sambrook); and Current Protocols in Molecular Biology, F.M. Ausubel et al., eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (1994 Supplement) (Ausubel). Product information from manufacturers of biological reagents and experimental equipment also provide information useful in known biological methods. Such manufacturers include the SIGMA chemical company (Saint Louis, MO), R&D systems (Minneapolis, MN), Pharmacia LKB Biotechnology (Piscataway, NJ), CLONTECH Laboratories, Inc. (Palo Alto, CA), Chem Genes Corp., Aldrich Chemical Company (Milwaukee, WI), Glen Research, Inc., GIBCO BRL Life Technologies, Inc. (Gaithersberg, MD), Fluka Chernica- Biochemika Analytika (Fluka Chemie AG, Buchs, Switzerland), Invitrogen, San Diego, CA, and Applied Biosystems (Foster City, CA), as well as many other commercial sources known to one of skill.

It will be appreciated that conservative substitutions of the given sequences can be used to produce nucleic acids which encode hyperrecombinogenic clones. "Conservatively modified variations" of a particular nucleic acid sequence refers to those nucleic acids which encode identical or essentially identical amino acid

sequences, or where the nucleic acid does not encode an amino acid sequence, to essentially identical sequences. Because of the degeneracy of the genetic code, a large number of functionally identical nucleic acids encode any given polypeptide. For instance, the codons CGU, CGC, CGA, CGG, AGA, and AGG all encode the amino acid arginine. Thus, at every position where an arginine is specified by a codon, the codon can be altered to any of the corresponding codons described without altering the encoded polypeptide. Such nucleic acid variations are "silent variations," which are one species of "conservatively modified variations." Every nucleic acid sequence herein which encodes a polypeptide also describes every possible silent variation. One of skill will recognize that each codon in a nucleic acid (except AUG, which is ordinarily the only codon, for methionine) can be modified to yield a functionally identical molecule by standard techniques. Accordingly, each "silent variation" of a nucleic acid which encodes a polypeptide is implicit in any described sequence. Furthermore, one of skill will recognize that individual substitutions, deletions or additions which alter, add or delete a single amino acid or a small percentage of amino acids (typically less than 5%, more typically less than 1%) in an encoded sequence are "conservatively modified variations" where the alterations result in the substitution of an amino acid with a chemically similar amino acid. Conservative substitution tables providing functionally similar amino acids are well known in the art. The following six groups each contain amino acids that are conservative substitutions for one another: 1) Alanine (A), Serine (S), Threonine (T); 2) Aspartic acid (D), Glutamic acid (E); 3) Asparagine (N), Glutamine (Q); 4) Arginine (R), Lysine (K); 5) Isoleucine (I), Leucine (L), Methionine (M), Valine (V); and 6) Phenylalanine (F), Tyrosine (Y), Tryptophan (W). See also, Creighton (1984) Proteins W.H. Freeman and Company. Finally, the addition of sequences which do not alter the encoded activity of a nucleic acid molecule, such as a non-functional sequence is a conservative modification of the basic nucleic acid.

One of skill will appreciate that many conservative variations of the nucleic acid constructs disclosed yield a functionally identical construct. For example, due to the degeneracy of the genetic code, "silent substitutions" (ie., substitutions of a nucleic

acid sequence which do not result in an alteration in an encoded polypeptide) are an implied feature of every nucleic acid sequence which encodes an amino acid. Similarly, "conservative amino acid substitutions," in one or a few amino acids in an amino acid sequence of a packaging or packageable construct are substituted with different amino acids with highly similar properties, are also readily identified as being highly similar to a disclosed construct. Such conservatively substituted variations of each explicitly disclosed sequence are a feature of the present invention.

Nucleic acids which hybridize under stringent conditions to the nucleic acids in the figures are a feature of the invention. "Stringent hybridization wash conditions" in the context of nucleic acid hybridization experiments such as Southern and northern hybridizations are sequence dependent, and are different under different environmental parameters. An extensive guide to the hybridization of nucleic acids is found in Tijssen (1993) *Laboratory Techniques in Biochemistry and Molecular Biology-Hybridization with Nucleic Acid Probes part I chapter 2 "overview of principles of hybridization and the strategy of nucleic acid probe assays"*, Elsevier, New York. Generally, highly stringent hybridization and wash conditions are selected to be about 5°C lower than the thermal melting point (T_m) for the specific sequence at a defined ionic strength and pH. The T_m is the temperature (under defined ionic strength and pH) at which 50% of the target sequence hybridizes to a perfectly matched probe. Very stringent conditions are selected to be equal to the T_m for a particular probe. In general, a signal to noise ratio of 2x (or higher) than that observed for an unrelated probe in the particular hybridization assay indicates detection of a specific hybridization.

Nucleic acids which do not hybridize to each other under stringent conditions are still substantially identical if the polypeptides which they encode are substantially identical. This occurs, e.g., when a copy of a nucleic acid is created using the maximum codon degeneracy permitted by the genetic code.

Finally, preferred nucleic acids encode hyper-recombinogenic RecA proteins which are at least one order of magnitude (10 times) as active as a wild-type RecA protein in a standard assay for Rec A activity.

4.6.1.2.2.14 RecE/RecT MEDIATED STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS IN VIVO

Like recA, recE and recT (or their homologues, for example the lambda recombination proteins red⁺ and red⁻) can stimulate homologous recombination in vivo. See, Muyrers et al. (1999) Nucleic Acids Res 27(6):1555-7 and Zhang et al. (1998) Nat Genet (2):123-8 Hyper-recombinogenic recE and recT are evolved by the same method as described for recA. Alternatively, variants with increased recombinogenicity are selected by their ability to cause recombination between a suicide vector (lacking an origin of replication) carrying a selectable marker, and a homologous region in either the chromosome or a stably-maintained episome.

A plasmid containing recA and recE genes is stochastic &/or non-stochastic mutagenized (either using these genes as single starting points, or by family stochastic &/or non-stochastic mutagenesis (with for example red⁺ and red⁻, or other homologous genes identified from available sequence databases). This stochastic &/or non-stochastic mutagenized library is then cloned into a vector with a selectable marker and transformed into an appropriate recombination-deficient strain. The library of cells would then be transformed with a second selectable marker, either borne on a suicide vector or as a linear DNA fragment with regions at its ends that are homologous to a target sequence (either in the plasmid or in the host chromosome). Integration of this marker by homologous recombination is a selectable event, dependent on the activity of the recE and recT gene products. The recE / recT genes are isolated from cells in which homologous recombination has occurred. The process is repeated several times to enrich for the most efficient variants before the next round of stochastic &/or non-stochastic mutagenesis is performed. In addition, cycles of recombination without selection can be performed to increase the diversity of a cell population prior to selection.

Once hyper-recombinogenic recE / recT genes are isolated they are used as described for hyper-recombinogenic recA. For example they are expressed (constitutively or conditionally) in a host cell to facilitate homologous recombination between variant gene fragments and homologues within the host cell. They are alternatively introduced by microinjection, biolistics, lipofection or other means into a host cell at the same time as the variant genes.

Hyper-recombinogenic recE/ recT (either of bacterial / phage origin, or from plant homologues) are useful for facilitating homologous recombination in plants. They are, for example, cloned into the Agrobacterium cloning vector, where they are expressed upon entry into the plant, thereby stimulating homologous recombination in the recipient cell.

In a preferred embodiment, recE/ recT are used and or generated in mutS strains.

4.6.1.2.2.15 MULTI-CYCLIC RECOMBINATION

As noted, protoplast fusion is an efficient means of recombining two microbial genomes. The process reproducibly results in about 10% of a non-selected population being recombinant chimeric organisms.

Protoplasts are cells that have been stripped of their cell walls by treatment in hypotonic medium with cell wall degrading enzymes. Protoplast fusion is the induced fusion of the membranes of two or more of these protoplasts by fusogenic agents such as polyethylene glycol. Fusion results in cytoplasmic mixing and places the genomes of the fused cells within the same membrane. Under these conditions recombination between the genomes is frequent.

The fused protoplasts are regenerated, and, during cell division, single genomes segregate into each daughter cell. Typically, 10% of these daughter cells have genomes that originate partially from more than one of the original parental protoplast genomes.

This result is similar to that of the crossing over of sister chromatids in eukaryotic cells during prophase of meiosis II. The percentage of daughter cells that are recombinant is just lower after protoplast fusion. While protoplast fusion does result in efficient recombination, the recombination predominantly occurs between two cells as in sexual recombination.

In order to efficiently generate libraries of whole genome stochastic &/or non-stochastic mutagenized libraries, daughter cells having genetic information originating from multiple parents are made.

In vitro DNA stochastic &/or non-stochastic mutagenesis results in the efficient poolwise recombination of multiple homologous DNA sequences. The stochastic &/or non-stochastic mutagenesis of full length genes from a mixed pool of small gene fragments requires multiple annealing and elongation cycles, the thermal cycles of the primerless PCR reaction. During each thermal cycle, many pairs of fragments anneal and are extended to form a combinatorial population of larger chimeric DNA fragments. After the first cycle of stochastic &/or non-stochastic mutagenesis, chimeric fragments contain sequences originating from two different parent genes. This is similar to the result of a single sexual cycle within a population, pairwise cross, or protoplast fusion. During the second cycle, these chimeric fragments can anneal with each other, or with other small fragments, resulting in chimeras originating from up to four different parental sequences.

This second cycle is analogous to the entire progeny from a single sexual cross inbreeding with itself. Further cycles will result in chimeras originating from 8, 16, 32, etc parental sequences and are analogous to further inbreedings of the progeny population. The power of in vitro DNA stochastic &/or non-stochastic mutagenesis is that a large combinatorial library can be generated from a single pool of DNA fragments stochastic &/or non-stochastic mutagenized by these recursive pairwise "matings." As described above, in vivo stochastic &/or non-stochastic mutagenesis strategies, such as protoplast fusion, result in a single pairwise mating reaction. Thus, to generate the level of diversity obtained by in vitro methods, in vivo methods are carried out recursively. That is, a pool of organisms is recombined and the progeny pooled, without selection, and then recombined again. This process is repeated for sufficient cycles to result in progeny having multiple parental sequences.

Described below is a method used to reassemble four strains of *Streptomyces coelicolor*. From the initial four strains each containing a unique nutritional marker, three to four rounds of recursive pooled protoplast fusion was sufficient to generate a population of stochastic &/or non-stochastic mutagenized organisms containing all 16 possible combinations of the four markers. This represents a 10^6 fold improvement in the generation of four parent progeny as compared to a single pooled fusion of the four strains.

Protoplasts were generated from several strains of *S. coelicolor*, pooled and fused. Mycelia were regenerated and allowed to sporulate. The spores were collected, allowed to grow into Mycelia, formed into protoplasts, pooled and fused and the process repeated for three to four rounds. the resulting spores were then subject to screening.

The basic protocol for generating a whole genome stochastic &/or non-stochastic mutagenized library from four *S. coelicolor* strains, each having one of four distinct

markers, was as follows. Four mycelial cultures, each of a strain having one of four different markers, were grown to early stationary phase. The mycelia from each were harvested by centrifugation and washed. Protoplasts from each culture were prepared as follows. Approximately 10^9 *S. coelicolor* spores were inoculated into 50ml YEME with 0.5% Glycine in a 250ml baffled flask. The spores were incubated at 30C for 36-40 hours in an orbital shaker. Mycelium were verified using a microscope. Some strains needed an additional day of growth. The culture was transferred into a 50ml tube and centrifuged at 4,000 rpm for 10 min. The mycelium were twice washed with 10.3% sucrose and centrifuged at 4,000 rpm for 10 min. (mycelium can be stored at about 80C after wash). 5ml of lysozyme was added to the about 0.5g of mycelium pellet. The pellet was suspended and incubated at 30C for 20-60 min., with gentle shaking every 10 min. The microscope was checked for protoplasting every 20 min. Once the majority were protoplasts, protoplasting was stopped by adding 10ml of P buffer. The protoplasts were filtered through cotton and the protoplast spun down at 3,000rpm for 7 min at room temperature. The supernatant was discarded and the protoplast gently resuspended, adding a suitable amount of P buffer according to the pellet size (usually about 50OW). Ten-fold serial dilutions were made in P buffer, and the protoplasts counted at a 10^{-2} dilution. Protoplasts were adjusted to 10^{10} protoplasts per ml.

The protoplasts from each culture were quantitated by microscopy. 10^8 protoplast from each culture were mixed in the same tube, washed, and then fused by the addition of 50% PEG. The fused protoplasts were diluted and plated regeneration medium and incubated until the colonies were sporulating (four days). Spores were harvested and washed. These spores represent a pool of all the recombinants and parents form the fusion.

A sample of the pooled spores was then used to inoculate a single liquid culture. The culture was grown to early stationary phase, the mycelia harvested, and protoplasts prepared. 10^8 protoplasts from this "mycelial library" were then fused with themselves

by the addition of 50%PEG. The protoplast fusion/regeneration/harvesting/protoplast preparation steps were repeated two times. The spores resulting from the fourth round of fusion were considered the "whole genome stochastic &/or non-stochastic mutagenized library" and they were screened for the frequency of the 16 possible combinations of the four markers

In particular, adding rounds of recombination prior to selection produced significant increases in the number of clones which incorporated all four of the relevant selectable markers, indicating that the population became increasingly diverse by recursive pooling and sporulation.

The four strains of the four parent stochastic &/or non-stochastic mutagenesis were each auxotrophic for three and prototrophic for one of four possible nutritional markers: arginine (A), cystine (C), proline (P), and/or uracil (U). Spores from each fusion were plated in each of the 16 possible combinations of these four nutrients, and the percent of the population growing on a particulate medium was calculated as the ratio of those colonies from a selective plate to those growing on a plate having all four nutrients (all variants grow on the medium having all four nutrients, thus the colonies from this plate thus represent the total viable population). The corrected percentages for each of the no, one, two, and three marker phenotypes were determined by subtracting the percentage of cells having additional markers that might grow on the medium having "unnecessary" nutrients. For example, the number of colonies growing on no additional nutrients (the prototroph) was subtracted from the number of colonies growing on any plate requiring nutrients.

4.6.1.2.16 WHOLE GENOME STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS THROUGH ORGANIZED HETERODUPLEX STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS

A new procedure to optimize phenotypes of interests by heteroduplex stochastic &/or

non-stochastic mutagenesis of cosmids libraries of the organism of choice, is provided. This procedure does not require protoplast fusion and is applicable to bacteria for which well-established genetic systems are available, including cosmid cloning, transformation, in vitro packaging/transfection and plasmid transfer/mobilization. Microorganism that can be improved by these methods include *Escherichia coli*, *Pseudomonas aeruginosa*, *Pseudomonas putida*, *Pseudomonas spp.*, *Rhizobium spp.*, *Xanthomonas spp.*, and other gram-negative organisms. This method is also applicable to Gram-positive microorganisms.

In step A, Chromosomal DNA of the organism to be improved is digested with suitable restriction enzymes and ligated into a cosmid. The cosmid used for cosmid-based heteroduplex guided whole genome stochastic &/or non-stochastic mutagenesis has at least two rare restriction enzyme recognition sites (e.g. Sfr and NotI) to be used for linearization in subsequent steps. Sufficient cosmids to represent the complete chromosome are purified and stored in 96-well microtiter dishes. In step B, small samples of the library are mutagenized in vitro using hydroxylamine or other mutagenic chemicals. In step C, a sample from each well of the mutagenized collection is used to transfet the target cells. In step D, the transfectants are assayed (as a pool from each mutagenized sample-well) for phenotypic improvements. Positives from this assay indicate that a cosmid from a particular well can confer phenotypic improvements and thus contain large genomic fragments that are suitable targets for heteroduplex mediated stochastic &/or non-stochastic mutagenesis. In step E, the transfected cells harboring a mutant library of the identified cosmid(s) are separated by plating on solid media and screened for independent mutants conferring an improved phenotype. In step F, DNA from positive cells is isolated and pooled by origin. In step G, the selected cosmid pools are divided so that one sample can be digested with Sfr and the other with NotI. These samples are pooled, denatured, reannealed, and religated.

In step K target cells are transfected with the resulting heteroduplexes and propagated

to allow "recombination" to occur between the strands of the heteroduplexes *in vivo*. The transfectants can be screened (the population will represent the pairwise recombinants) or, commonly, as represented by step 1, the recombined cosmids are further stochastic &/or non-stochastic mutagenized by recursive *in vitro* heteroduplex formation and *in vivo* recombination (to generate a complete combinatorial library of the possible mutations) prior to screening. An additional mutagenesis step could also be added for increased diversity during the stochastic &/or non-stochastic mutagenesis process.

In step J, once several cosmids harboring different distributed loci have been improved, they are combined into the same host by chromosome integration. This organism can be used directly or subjected to a new round of heteroduplex guided whole genome stochastic &/or non-stochastic mutagenesis.

4.7. SPECIALIZED METHODS

4.7.1 TARGETED STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS- HOT SPOTS

In one aspect, targeted homologous genes are cloned into specific regions of the genome (e.g., by homologous recombination or other targeting procedures) which are known to be recombination "hot spots" (i.e., regions showing elevated levels of recombination compared to the average level of recombination observed across an entire genome), or known to be proximal to such hot spots. The resulting recombinant strains are mated recursively. During meiotic recombination, homologous recombinant genes recombine, thereby increasing the diversity of the genes. After several cycles of recombination by recursive mating, the resulting cells are screened.

4.7.2 STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS USING YEASTS

Yeast are subspecies of fungi that grow as single cells. Yeasts are used for the production of fermented beverages and leavening, for production of ethanol as a fuel,

low molecular weight compounds, and for the heterologous production of proteins and enzymes (see accompanying list of yeast strains and their uses). Commonly used strains of yeast include *Saccharomyces cerevisiae*, *Pichia* sp., *Canidio* sp. and *Schizosaccharomyces pombe*.

Several types of vectors are available for cloning in yeast including integrative plasmid (YIp), yeast replicating plasmid (YRp, such as the 2 circle based vectors), yeast episomal plasmid (YEp), yeast centromeric plasmid (YCP), or yeast artificial chromosome (YAC). Each vector can carry markers useful to select for the presence of the plasmid such as LUE2, URA3, and HIS3, or the absence of the plasmid such as URA3 (a gene that is toxic to cells grown in the presence of 5-fluoro orotic acid).

Many yeasts have a sexual cycle and asexual (vegetative) cycles. The sexual cycle involves the recombination of the whole genome of the organism each time the cell passes through meiosis. For example, when diploid cells of *S. cerevisiae* are exposed to nitrogen and carbon limiting conditions, diploid cells undergo meiosis to form ascii. Each ascus holds four haploid spores, two of mating type "a" and two of mating type "α". Upon return to rich medium, haploid spores of opposite mating type mate to form diploid cells once again. Asiospores of opposite mating type can mate within the ascus, or if the ascus is degraded, for example with zymolase, the haploid cells are liberated and can mate with spores from other ascii. This sexual cycle provides a format to reassemble endogenous genomes of yeast and/or exogenous fragment libraries inserted into yeast vectors. This process results in swapping or accumulation of hybrid genes, and for the stochastic &/or non-stochastic mutagenesis of homologous sequences shared by mating cells.

Yeast strains having mutations in several known genes have properties useful for stochastic &/or non-stochastic mutagenesis. These properties include increasing the frequency of recombination and increasing the frequency of spontaneous mutations

within a cell. These properties can be the result of mutation of a coding sequence or altered expression (usually overexpression) of a wildtype coding sequence. The HO nuclease effects the transposition of HML α / and HMR α / to the MAT locus resulting in mating type switching. Mutants in the gene encoding this enzyme do not switch their mating type and can be employed to force crossing between strains of defined genotype, such as ones that harbor a library or have a desired phenotype and to prevent in breeding of starter strains. PMS1, MLH1, MSH2, MSH6 are involved in mismatch repair. Mutations in these genes all have a mutator phenotype (Chambers et al., Mol. Cell. Biol. 16, 6110-6120 (1996)). Mutations in TOP3 DNA topoisomerase have a 6-fold enhancement of interchromosomal homologous recombination (Bailis et al., Molecular and Cellular Biology 12, 4988-4993 (1992)). The RAD50-57 genes confer resistance to radiation. Rad3 functions in excision of pyrimidine dimers. RAD52 functions in gene conversion. RAD50, MRE11, XRS2 function in both homologous recombination and illegitimate recombination. HOP1, RED1 function in early meiotic recombination (Mao-Draayer, Genetics 144, 71-86) Mutations in either HOP1 or RED 1 reduce double stranded breaks at the HIS2 recombination hotspot. Strains deficient in these genes are useful for maintaining stability in hyper recombinogenic constructs such as tandem expression libraries carried on YACs. Mutations in HPR1 are hyperrecombinogenic. HDF1 has DNA end binding activity and is involved in double stranded break repair and V(D)J recombination.

Strains bearing this mutation are useful for transformation with random genomic fragments by either protoplast fusion or electroporation. Kar-1 is a dominant mutation that prevents karyogamy. Kar-1 mutants are useful for the directed transfer of single chromosomes from a donor to a recipient strain. This technique has been widely used in the transfer of YACs between strains, and is also useful in the transfer of evolved genes/chromosomes to other organisms (Markie, YA C Protocols, (Humana Press, Totowa, NJ, 1996)). HOT1 is an *S. cerevisiae* recombination hotspot within the promoter and enhancer region of the rDNA repeat sequences. This locus induces mitotic recombination at adjacent sequences- presumably due to its high level transcription. Genes and/or pathways inserted under the transcriptional control of this

region undergo increased mitotic recombination. The regions surrounding the arg 4 and his 4 genes are also recombination hot spots, and genes cloned in these regions have an increased probability of undergoing recombination during meiosis.

Homologous genes can be cloned in these regions and stochastic &/or non-stochastic mutagenized in vivo by recursively mating the recombinant strains. CDC2 encodes polymerase and is necessary for mitotic gene conversion. Overexpression of this gene can be used in a reassembler or mutator strain. A temperature sensitive mutation in CDC4 halts the cell cycle at G1 at the restrictive temperature and could be used to synchronize protoplasts for optimized fusion and subsequent recombination.

As with filamentous fungi, the general goals of stochastic &/or non-stochastic mutagenesis yeast include improvement in yeast as a host organism for genetic manipulation, and as a production apparatus for various compounds. One desired property in either case is to improve the capacity of yeast to express and secrete a heterologous protein. The following example describes the use of stochastic &/or non-stochastic mutagenesis to evolve yeast to express and secrete increased amounts of RNase A.

RNase A catalyzes the cleavage of the P-0_{5'} bond of RNA specifically after pyrimidine nucleotides. The enzyme is a basic 124 amino acid polypeptide that has 8 half cystine residues, each required for catalysis. YEpWL-RNase A is a vector that effects the expression and secretion of RNaseA from the yeast *S. cerevisiae*, and yeast harboring this vector secrete 1-2 mg of recombinant RNase A per liter of culture medium (del Cardayre et al., Protein Engineering 8(3):26, 1-273 (1995)). This overall yield is poor for a protein heterologously expressed in yeast and can be improved at least 10-100 fold by stochastic &/or non-stochastic mutagenesis. The expression of RNaseA is easily detected by several plate and microtitre plate assays (del Cardayre & Raines, Biochemistry 33, 6031-6037 1994)). Each of the described formats for

whole genome stochastic &/or non-stochastic mutagenesis can be used to reassemble a strain of *S. cerevisiae* harboring YepWL-RNase A, and the resulting cells can be screened for the increased secretion of RNase A into the medium. The new strains are cycled recursively through the stochastic &/or non-stochastic mutagenesis format, until sufficiently high levels of RNase A secretion is observed. The use of RNase A is particularly useful since it not only requires proper folding and disulfide bond formation but also proper glycosylation. Thus numerous components of the expression, folding, and secretion systems can be optimized. The resulting strain is also evolved for improved secretion of other heterologous proteins.

4.7.3 REASSEMBLE TO INCREASE TOLERANCE OF YEAST TO ETHANOL

Another goal of stochastic &/or non-stochastic mutagenesis yeast is to increase the tolerance of yeast to ethanol. Such is useful both for the commercial production of ethanol, and for the production of more alcoholic beers and wines. The yeast strain to be stochastic &/or non-stochastic mutagenized acquires genetic material by exchange or transformation with other strain(s) of yeast, which may or may not be known to have superior resistance to ethanol. The strain to be evolved is stochastic &/or non-stochastic mutagenized and shufflants are selected for capacity to survive exposure to ethanol. Increasing concentrations of ethanol can be used in successive rounds of stochastic &/or non-stochastic mutagenesis. The same principles can be used to reassemble baking yeasts for improved osmotolerance.

4.7.4 CAPACITY TO GROW UNDER DESIRED NUTRITIONAL CONDITIONS

Another desired property of stochastic &/or non-stochastic mutagenesis yeast is capacity to grow under desired nutritional conditions. For example, it is useful to yeast to grow on cheap carbon sources such as methanol, starch, molasses, cellulose, cellobiose, or xylose depending on availability. The principles of stochastic &/or non-

stochastic mutagenesis and selection are similar to those discussed for filamentous fungi.

4.7.5 TO PRODUCE SECONDARY METABOLITES

Another desired property is capacity to produce secondary metabolites naturally produced by filamentous fungi or bacteria. Examples of such secondary metabolites are cyclosporin A, taxol, and cephalosporins. The yeast to be evolved undergoes genetic exchange or is transformed with DNA from organism(s) that produce the secondary metabolite. For example, fungi producing taxol include *Taxomyces andreanae* and *Pestalotopsis microspora* (Stierle et al., *Science* 260, 214-216 (1993); Strobel et al., *Microbiol.* 142, 435440 (1996)). DNA can also be obtained from trees that naturally produce taxol, such as *Taxus brevifolia*. DNA encoding one enzyme in the taxol pathway, taxadiene synthase, which it is believed catalyzes the committed step in taxol biosynthesis and may be rate limiting in overall taxol production, has been cloned (Wildung & Croteau, *J Biol Chem.* 271, 9201-4 (1996)). The DNA is then stochastic &/or non-stochastic mutagenized, and shufflants are screened/selected for production of the secondary metabolite. For example, taxol production can be monitored using antibodies to taxol, by mass spectroscopy or UV spectrophotometry. Alternatively, production of intermediates in taxol synthesis or enzymes in the taxol synthetic pathway can be monitored. Concetti & Ripani, *Biol Chem. Hoppe Seyler* 375, 419-23 (1994). Other examples of secondary metabolites are polyols, amino acids, polyketides, non-ribosomal polypeptides, ergosterol, carotenoids, terpenoids, sterols, vitamin E, and the like.

4.7.6 INCREASE ABILITY TO SEPARATE IN ETHANOL

Another desired property is to increase the flocculence of yeast to facilitate separation in preparation of ethanol. Yeast can be stochastic &/or non-stochastic mutagenized by any of the procedures noted above with selection for stochastic &/or non-stochastic mutagenized yeast forming the largest clumps.

4.7.6.1 EXEMPLARY PROCEDURE FOR YEAST PROTOPLASTING

Protoplast preparation in yeast is reviewed by Morgan, in Protoplasts (Birkhauser Verlag, Basel, 1983). Fresh cells ($\sim 10^8$) are washed with buffer, for example 0.1 M potassium phosphate, then resuspended in this same buffer containing a reducing agent, such as 50 mM DTT, incubated for 1 h at 30°C with gentle agitation, and then washed again with buffer to remove the reducing agent. These cells are then resuspended in buffer containing a cell wall degrading enzyme, such as Novozyme 234 (1 mg/mL), and any of a variety of osmotic stabilizers, such as sucrose, sorbitol, NaCl, KCl, MgSO₄, MgCl₂, or NH₄Cl at any of a variety of concentrations. These suspensions are then incubated at 30°C with gentle shaking (~60 rpm) until protoplasts are released. To generate protoplasts that are more likely to produce productive fusants several strategies are possible.

Protoplast formation can be increased if the cell cycle of the protoplasts have been synchronized to be halted at G1. In the case of *S. cerevisiae* this can be accomplished by the addition of mating factors, either a or alpha (Curran & Carter, J Gen. Microbiol. 129, 1589-1591 (1983)). These peptides act as adenylate cyclase inhibitors which by decreasing the cellular level of cAMP arrest the cell cycle at G1. In addition, sex factors have been shown to induce the weakening of the cell wall in preparation for the sexual fusion of a and alpha cells (Crandall & Brock, Bacteriol. Rev. 32, 139-163 (1968); Osumi et al., Arch. Microbiol 97, 27-38 (1974)). Thus in the preparation of protoplasts, cells can be treated with mating factors or other known inhibitors of adenylate cyclase, such as leflunomide or the killer toxin from *K. lactis*, to arrest them at G1 (Sugisaki et al., Nature 304, 464- 466 (1983)). Then after fusing of the protoplasts (step 2), cAMP can be added to the regeneration medium to induce S-phase and DNA synthesis. Alternatively, yeast strains having a temperature sensitive mutation in the CDC4 gene can be used, such that cells could be synchronized and arrested at G1. After fusion cells are returned to the permissive temperature so that DNA synthesis and growth resumes.

Once suitable protoplasts have been prepared, it is necessary to induce fusion by physical or chemical means. An equal number of protoplasts of each cell type is mixed in phosphate buffer (0.2 M, pH 5.8, 2×10^8 cells/mL) containing an osmotic stabilizer, for example 0.8 M NaCl, and PEG 6000 (33% w/v) and then incubated at 30°C for 5 min while fusion occurs. Polyols, or other compounds that bind water, can be employed. The fusants are then washed and resuspended in the osmotically stabilized buffer lacking PEG, and transferred to osmotically stabilized regeneration medium on/in which the cells can be selected or screened for a desired property.

4.7.7 STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS USING ARTIFICIAL CHROMOSOMES

Yeast artificial chromosomes (Yacs) are yeast vectors into which very large DNA fragments (e.g., 50-2000 kb) can be cloned (see, e.g., Monaco & Larin, Trends. Biotech. 12(7), 280-286 (1994); Ramsay, Mol Biotechnol 1(2), 181-201 1994; Huxley, Genet. Eng. 16, 65-91 (1994); Jakobovits, Curr. Biol. 4(8), 761-3 (1994); Lamb & Gearhart, Curr. Opin. Genet. Dev. 5(3), 342-8 (1995); Montoliu et al., Reprod Fertil. Dev. 6, 577-84 (1994)). These vectors have telomeres (Tel), a centromere (Cen), an autonomously replicating sequence (ARS), and can have genes for positive (e.g., TRPI) and negative (e.g., URA3) selection. YACs are maintained, replicated, and segregate as other yeast chromosomes through both meiosis and mitosis thereby providing a means to expose cloned DNA to true meiotic recombination.

YACs provide a vehicle for the stochastic &/or non-stochastic mutagenesis of libraries of large DNA fragments *in vivo*. The substrates for stochastic &/or non-stochastic mutagenesis are typically large fragments from 20 kb to 2 Mb. The fragments can be random fragments or can be fragments known to encode a desirable property. For example, a fragment might include an operon of genes involved in

production of antibiotics. Libraries can also include whole genomes or chromosomes. Viral genomes and some bacterial genomes can be cloned intact into a single YAC. In some libraries, fragments are obtained from a single organism. Other libraries include fragment variants, as where some libraries are obtained from different individuals or species. Fragment variants can also be generated by induced mutation. Typically, genes within fragments are expressed from naturally associated regulatory sequences within yeast. However, alternatively, individual genes can be linked to yeast regulatory elements to form an expression cassette, and a concatemer of such cassettes, each containing a different gene, can be inserted into a YAC.

In some instances, fragments are incorporated into the yeast genome, and stochastic &/or non-stochastic mutagenesis is used to evolve improved yeast strains. In other instances, fragments remain as components of YACs throughout the stochastic &/or non-stochastic mutagenesis process, and after acquisition of a desired property, the YACs are transferred to a desired recipient cell.

4.7.8 STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS OF GENES FOR BIOREMEDIALTION

Modern industry generates many pollutants for which the environment can no longer be considered an infinite sink. Naturally occurring microorganisms are able to metabolize thousands of organic compounds, including many not found in nature (e.g xenobiotics). Bioremediation, the deliberate use of microorganisms for the biodegradation of man-made wastes, is an emerging technology that offers cost and practicality advantages over traditional methods of disposal. The success of bioremediation depends on the availability of organisms that are able to detoxify or mineralize pollutants.

Microorganisms capable of degrading specific pollutants can be generated by genetic engineering and recursive sequence recombination. Although bioremediation is an aspect of pollution control, a more useful approach in the long term is one of

prevention before industrial waste is pumped into the environment. Exposure of industrial waste streams to recursive sequence recombination-generated microorganisms capable of degrading the pollutants they contain would result in detoxification or mineralization of these pollutants before the waste stream enters the environment. Issues of releasing recombinant organisms can be avoided by containing them within bioreactors fitted to the industrial effluent pipes. This approach would also allow the microbial mixture used to be adjusted to best degrade the particular wastes being produced. Finally, this method would avoid the problems of adapting to the outside world and dealing with competition that face many laboratory microorganisms.

In the wild, microorganisms have evolved new catabolic activities enabling them to exploit pollutants as nutrient sources for which there is no competition. However, pollutants that are present at low concentrations in the environment may not provide a sufficient advantage to stimulate the evolution of catabolic enzymes. For a review of such naturally occurring evolution of biodegradative pathways and the manipulation of some of microorganisms by classical techniques, see Ramos et al., Bio/Technology "12:1349-1355 (1994).

Generation of new catabolic enzymes or pathways for bioremediation has thus relied upon deliberate transfer of specific genes between organisms (Wackett et al., *supra*), forced matings between bacteria with specific catabolic capabilities (Brenner et al. Biodegradation 5:359-377 (1994)), or prolonged selection in a chemostat. Some researchers have attempted to facilitate evolution via naturally occurring genetic mechanisms in their chemostat selections by including microorganisms with a variety of catabolic pathways (Kellogg et. al. Science 214:1133-1135 (1981); Chakrabarty American Society of Micro. Biol. News 62:130-137 (1996)). For a review of efforts in this area, see Cameron et al. Applied Biochem. Biotech 38:105-140 (1993).

Current efforts in improving organisms for bioremediation take a labor-intensive approach in which many parameters are optimized independently, including transcription efficiency from native and heterologous promoters, regulatory circuits

and translational efficiency as well as improvement of protein stability and activity (Timmis et al. Ann. Rev. Microbiol. 48:525- 527 (1994)).

A recursive sequence recombination approach overcomes a number of limitations in the bioremediation capabilities of naturally occurring microorganisms. Both enzyme activity and specificity can be altered, simultaneously or sequentially, by the methods of the invention. For example, catabolic enzymes can be evolved to increase the rate at which they act on a substrate. Although knowledge of a rate-limiting step in a metabolic pathway is not required to practice the invention, rate- limiting proteins in pathways can be evolved to have increased expression and/or activity, the requirement for inducing substances can be eliminated, and enzymes can be evolved that catalyze novel reactions.

Some examples of chemical targets for bioremediation include but are not limited to benzene, xylene, and toluene, camphor, naphthalene, halogenated hydrocarbons, polychlorinated biphenyls (PCBs), trichlorethylene, pesticides such as pentachlorophenyls (PCPs), and herbicides such as atrazine.

4.7.8.1 AROMATIC HYDROCARBONS

Preferably, when an enzyme is "evolved" to have a new catalytic function, that function is expressed, either constitutively or in response to the new substrate. Recursive sequence recombination subjects both structural and regulatory elements (including the structure of regulatory proteins) of a protein to recombinogenic mutagenesis simultaneously. Selection of mutants that are efficiently able to use the new substrate as a nutrient source will be sufficient to ensure that both the enzyme and its regulation are optimized, without detailed analysis of either protein structure or operon regulation.

Examples of aromatic hydrocarbons include but are not limited to benzene, xylene, toluene, biphenyl, and polycyclic aromatic hydrocarbons such as pyrene and naphthalene. These compounds are metabolized via catechol intermediates.

Degradation of catechol by *Pseudomonas putida* requires induction of the catabolic operon by cis, cis-muconate which acts on the CatR regulatory protein. The binding site for the CatR protein is G-N₁₁-A, while the optimal sequence for the LysR class of activators (of which CatR is a member) is T-N₁₁-A. Mutation of the G to a T in the CatR binding site enhances the expression of catechol metabolizing genes (Chakrabarty, American Society of Microbiology News 62:130-137 (1996)). This demonstrates that the control of existing catabolic pathways is not optimized for the metabolism of specific xenobiotics. It is also an example of a type of mutant that would be expected from recursive sequence recombination of the operon followed by selection of bacteria that are better able to degrade the target compound.

As an example of starting materials, dioxygenases are required for many pathways in which aromatic compounds are catabolized. Even small differences in dioxygenase sequence can lead to significant differences in substrate specificity (Furukawa et al. J. Bact. 175:5224-5232 (1993); Erickson et al. App. Environ. Micro. 59:3858-3862 (1993)). A hybrid enzyme made using sequences derived from two "parental" enzymes may possess catalytic activities that are intermediate between the parents (Erickson, *ibid.*), or may actually be better than either parent for a specific reaction (Furukawa et al. J. Bact. 176:2121-2123 (1994)). In one of these cases site directed mutagenesis was used to generate a single polypeptide with hybrid sequence (Erickson, *ibid.*); in the other, a four subunit enzyme was produced by expressing two subunits from each of two different dioxygenases (Furukawa, *ibid.*). Thus, sequences from one or more genes encoding dioxygenases can be used in the recursive sequence recombination techniques of the instant invention, to generate enzymes with new specificities. In addition, other features of the catabolic pathway can also be evolved using these techniques, simultaneously or sequentially, to optimize the metabolic pathway for an activity of interest.

4.7.8.2 HALOGENATED HYDROCARBONS

Large quantities of halogenated hydrocarbons are produced annually for uses as solvents and biocides. These include, in the United States alone, over 5 million tons of

both 1,2-dichloroethane and vinyl chloride used in PVC production in the U.S. alone. The compounds are largely not biodegradable by processes in single organisms, although in principle haloaromatic catabolic pathways can be constructed by combining genes from different microorganisms. Enzymes can be manipulated to change their substrate specificities. Recursive sequence recombination offers the possibility of tailoring enzyme specificity to new substrates without needing detailed structural analysis of the enzymes.

As an example of possible starting materials for the methods of the instant invention, Wackett et al. (*Nature* 368:627-629 (1994)) recently demonstrated that through classical techniques a recombinant *Pseudomonas* strain in which seven genes encoding two multi-component oxygenases are combined, generated a single host that can metabolize polyhalogenated compounds by sequential reductive and oxidative techniques to yield non-toxic products. These and/or related materials can be subjected to the techniques discussed above so as to evolve and optimize a biodegradative pathway in a single organism.

Trichloroethylene is a significant groundwater contaminant. It is degraded by microorganisms in a cometabolic way (i.e., no energy or nutrients are derived). The enzyme must be induced by a different compound (e.g., *Pseudomonas cepacia* uses toluene-4-monoxygenase, which requires induction by toluene, to destroy trichloroethylene). Furthermore, the degradation pathway involves formation of highly reactive epoxides that can inactivate the enzyme (Timmis et al. *Ann. Rev. Microbiol.* 48:525-557 (1994)). The recursive sequence recombination techniques of the invention could be used to mutate the enzyme and its regulatory region such that it is produced constitutively, and is less susceptible to epoxide inactivation. In some embodiments of the invention, selection of hosts constitutively producing the enzyme and less susceptible to the epoxides can be accomplished by demanding growth in the presence of increasing concentrations of trichloroethylene in the absence of inducing substances.

4.7.8.3 POLYCHLORINATED BIPHENYLS AND POLYCYCLIC AROMATIC HYDROCARBONS

Polychlorinated Biphenyls (PCBs) and Polycyclic Aromatic Hydrocarbons (PAHs) PCBs and PAHs are families of structurally related compounds that are major pollutants at many Superfund sites. Bacteria transformed with plasmids encoding enzymes with broader substrate specificity have been used commercially. In nature, no known pathways have been generated in a single host that degrade the larger PAHs or more heavily chlorinated PCBs. Indeed, often the collaboration of anaerobic and aerobic bacteria are required for complete metabolism.

Thus, likely sources for starting material for recursive sequence recombination include identified genes encoding PAH-degrading catabolic pathways on large (20-10OKB) plasmids (Sanseverino et al. *Applied Environ. Micro.* 59:1931-1937 (1993); Simon et al. *Gene* 127:31-37 (1993); Zylstra et al. *Annals of the NY Acad. Sci.* 721:386-398 (1994)); while biphenyl and PCB-metabolizing enzymes are encoded by chromosomal gene clusters, and in a number of cases have been cloned onto plasmids (Hayase et al. *J. Bacteriol.* 172:1160-1164 (1990); Furukawa et al. *Gene* 98:21-28 (1992); Hofer et al. *Gene* 144:9-16 (1994)). The materials can be subjected to the techniques discussed above so as to evolve a biodegradative pathway in a single organism.

Substrate specificity in the PCB pathway largely results from enzymes involved in initial dioxygenation reactions, and can be significantly altered by mutations in those enzymes (Erickson et al. *Applied Environ. Micro.* 59:3858-38662 (1993); Furukawa et al. *J. Bact.* 175:5224-5232 (1993). Mineralization of PAHs and PCBs requires that the downstream pathway is able to metabolize the products of the initial reaction (Brenner et al. *Biodegradation* 5:359-377 (1994)). In this case, recursive sequence recombination of the entire pathway with selection for bacteria able to use the PCE or PAH as the sole carbon source will allow production of novel PCB and PAH degrading bacteria.

4.7.8.4 HERBICIDES

A general method for evolving genes for the catabolism of insoluble herbicides is exemplified as follows for atrazine. Atrazine [2-chloro-4-(ethylamino)- 6-(isopropylamino)-1,3,5-triazine] is a moderately persistent herbicide which is frequently detected in ground and surface water at concentrations exceeding the 3 ppb health advisory level set by the EPA. Atrazine can be slowly metabolized by a *Pseudomonas* species (Mandelbaum et al. *Appl. Environ. Micro.* 61:1451-1457 (1995)). The enzymes catalyzing the first two steps in atrazine metabolism by *Pseudomonas* are encoded by genes AtzA and AtzB (de Souza et al. *Appl. Environ. Micro.* 61:3373-3378 (1995)). These genes have been cloned in a 6.8 kb fragment into pUC18 (AtzAB-pUC). *E. coli* carrying this plasmid converts atrazine to much more soluble metabolites. It is thus possible to screen for enzyme activity by growing bacteria on plates containing atrazine. The herbicide forms an opaque precipitate in the plates, but cells containing AtzAB-pU18 secrete atrazine degrading enzymes, leading to a clear halo around those cells or colonies. Typically, the size of the halo and the rate of its formation can be used to assess the level of activity so that picking colonies with the largest halos allows selection of the more active or highly produced atrazine degrading enzymes.

Thus, the plasmids carrying these genes can be subjected to the recursive sequence recombination formats described above to optimize the catabolism of atrazine in *E. coli* or another host of choice, including *Pseudomonas*. After each round of recombination, screening of host colonies expressing the evolved genes can be done on agar plates containing atrazine to observe halo formation. This is a generally applicable method for screening enzymes that metabolize insoluble compounds to those that are soluble (e.g., polycyclic aromatic hydrocarbons). Additionally, catabolism of atrazine can provide a source of nitrogen for the cell; if no other nitrogen is available, cell growth will be limited by the rate at which the cells can catabolize nitrogen. Cells able to utilize atrazine as a nitrogen source can thus be selected from a background of non-utilizers or poor-utilizers.

4.7.8.5 HEAVY METAL DETOXIFICATION

Bacteria are used commercially to detoxify arsenate waste generated by the mining of arsenopyrite gold ores. As well as mining effluent, industrial waste water is often contaminated with heavy metals (e.g., those used in the manufacture of electronic components and plastics). Thus, simply to be able to perform other bioremedial functions, microorganisms must be resistant to the levels of heavy metals present, including mercury, arsenate, chromate, cadmium, silver, etc.

A strong selective pressure is the ability to metabolize a toxic compound to one less toxic. Heavy metals are toxic largely by virtue of their ability to denature proteins (Ford et al. Bioextraction and Biodeterioration of Metals, p. 1-23). Detoxification of heavy metal contamination can be effected in a number of ways including changing the solubility or bioavailability of the metal, changing its redox state (e.g. toxic mercuric chloride is detoxified by reduction to the much more volatile elemental mercury) and even by bioaccumulation of the metal by immobilized bacteria or is plants. The accumulation of metals to a sufficiently high concentration allows metal to be recycled; smelting burns off the organic part of the organism, leaving behind reusable accumulated metal. Resistances to a number of heavy metals (arsenate, cadmium, cobalt, chromium, copper, mercury, nickel, lead, silver, and zinc) are plasmid encoded in a number of species including *Staphylococcus* and *Pseudomonas* (Silver et al. Environ. Health Perspect. 102:107-113 (1994); Ji et al. J. Ind. Micro. 14:61-75 (1995)). These genes also confer heavy metal resistance on other species as well (e.g., *E. coli*). The recursive sequence recombination techniques of the instant invention (RSR) can be used to increase microbial heavy metal tolerances, as well as to increase the extent to which cells will accumulate heavy metals. For example, the ability of *E. coli* to detoxify arsenate can be improved at least 100-fold by RSR.

Cyanide is very efficiently used to extract gold from rock containing as little as 0.2 oz per ton. This cyanide can be microbially neutralized and used as a nitrogen source by fungi or bacteria such as *Pseudomonas fluorescens*. A problem with microbial cyanide degradation is the presence of toxic heavy metals in the leachate. RSR can be used to increase the resistance of bioremedial microorganisms to toxic heavy metals, so that they will be able to survive the levels present in many industrial and Superfund sites. This will allow them to biodegrade organic pollutants including but not limited to

aromatic hydrocarbons, halogenated hydrocarbons, and biocides.

4.7.8.6 MICROBIAL MINING

"Bioleaching" is the process by which microbes convert insoluble metal deposits (usually metal sulfides or oxides) into soluble metal sulfates. Bioleaching is commercially important in the mining of arsenopyrite, but has additional potential in the detoxification and recovery of metals and acids from waste dumps. Naturally occurring bacteria capable of bioleaching are reviewed by Rawlings and Silver (Bio/Technology 13:773-778 (1995)).

These bacteria are typically divided into groups by their preferred temperatures for growth. The more important mesophiles are *Thiobacillus* and *Leptospirillum* species. Moderate thermophiles include *Sulfbacillus* species. Extreme thermophiles include *Sulfolobus* species. Many of these organisms are difficult to grow in commercial industrial settings, making their catabolic abilities attractive candidates for transfer to and optimization in other organisms such as *Pseudomonas*, *Rhodococcus*, *T. ferrooxidans* or *E. coli*. Genetic systems are available for at least one strain of *T. ferrooxidans*, allowing the manipulation of its genetic material on plasmids.

The recursive sequence recombination methods described above can be used to optimize the catalytic abilities in native hosts or heterologous hosts for evolved bioleaching genes or pathways, such as the ability to convert metals from insoluble to soluble salts. In addition, leach rates of particular ores can be improved as a result of, for example, increased resistance to toxic compounds in the ore concentrate, increased specificity for certain substrates, ability to use different substrates as nutrient sources, and so on.

4.7.8.7 OIL DESULFURIZATION

The presence of sulfur in fossil fuels has been correlated with corrosion of pipelines,

pumping, and refining equipment, and with the premature breakdown of combustion engines. Sulfur also poisons many catalysts used in the refining of fossil fuels. The atmospheric emission of sulfur combustion products is known as acid rain.

Microbial desulfurization is an appealing bioremediation application. Several bacteria have been reported that are capable of catabolizing dibenzothiophene (DBT), which is the representative compound of the class of sulfur compounds found in fossil fuels. U.S. Patent No. 5,356,801 discloses the cloning of a DNA molecule from *Rhodococcus rhodochrous* capable of biocatalyzing the desulfurization of oil. Denome et al. (Gene 175:6890-6901 (1995)) disclose the cloning of a 9.8 kb DNA fragment from *Pseudomonas* encoding the upper naphthalene catabolizing pathway which also degrades dibenzothiophene. Other genes have been identified that perform similar functions (disclosed in U.S. 5,356,801).

The activity of these enzymes is currently too low to be commercially viable, but the pathway could be increased in efficiency using the recursive sequence recombination techniques of the invention. The desired property of the genes of interest is their ability to desulfurize dibenzothiophene or its alkyl or aryl substituted analogues. In some embodiments of the invention, selection is preferably accomplished by coupling this pathway to one providing a nutrient to the bacteria. Thus, for example, desulfurization of dibenzothiophene results in formation of hydroxybiphenyl.

This is a substrate for the biphenyl-catabolizing pathway which provides carbon and energy. Selection would thus be done by "stochastic &/or non-stochastic mutagenesis" the dibenzothiophene genes and transforming them into a host containing the biphenyl- catabolizing pathway. Increased dibenzothiophene desulfurization will result in increased nutrient availability and increased growth rate. Once the genes have been evolved they are easily separated from the biphenyl degrading genes. The latter are undesirable in the final product since the object is to desulfurize without decreasing the energy content of the oil. Alkyl or aryl substituted dibenzothiophenes can be detected by changes in fluorescence (Krawiec, S., Devel. Indus. Microbiology 31:103-114 (1990)) or by detection of phenol groups formed as a result of desulfurization (Dacre, J.C. Anal. Chem. 43:589-591 (1971)).

4.7.8.8 ORGANO-NITRO COMPOUNDS

Organo-nitro compounds are used as explosives, dyes, drugs, polymers and antimicrobial agents. Biodegradation of these compounds occurs usually by way of reduction of the nitrate group, catalyzed by nitroreductases, a family of broadly-specific enzymes. Partial reduction of organo-nitro compounds often results in the formation of a compound more toxic than the original (Hassan et al. 1979 Arch Bioch Biop. 196:385- 395). Recursive sequence recombination of nitroreductases can produce enzymes that are more specific, and able to more completely reduce (and thus detoxify) their target compounds (examples of which include but are not limited to nitrotoluenes and nitrobenzenes). Nitro-reductases can be isolated from bacteria isolated from explosive-contaminated soils, such as *Morganella morganii* and *Enterobacter cloacae* (Bryant et. al., 1991. J. Biol Chem. 266:4126-4130). A preferred selection method is to look for increased resistance to the organo-nitro compound of interest, since that will indicate that the enzyme is also able to reduce any toxic partial reduction products of the original compound.

4.7.8.9 ALTERNATIVE SUBSTRATES FOR CHEMICAL SYNTHESIS

Metabolic engineering can be used to alter microorganisms that produce industrially useful chemicals, so that they will grow using alternate and more abundant sources of nutrients, including human-produced industrial wastes. This typically involves providing both a transport system to get the alternative substrate into the engineered cells and catabolic enzymes from the natural host organisms to the engineered cells.

In some instances, enzymes can be secreted into the medium by engineered cells to degrade the alternate substrate into a form that can more readily be taken up by the engineered cells; in other instances, a batch of engineered cells can be grown on one preferred substrate, then lysed to liberate hydrolytic enzymes for the alternate substrate into the medium, while a second inoculum of the same engineered host or a second host is added to utilize the hydrolyzate.

The starting materials for recursive sequence recombination will typically be genes for utilization of a substrate or its transport. Examples of nutrient sources of interest include but are not limited to lactose, whey, galactose, mannitol, xylan, cellobiose, cellulose and sucrose, thus allowing cheaper production of compounds including but not limited to ethanol, tryptophan, rhamnolipid surfactants, xanthan gum, and polyhydroxylalkanoate. For a review of such substrates as desired target substances, see Cameron et al. (*Appl. Biochem. Biotechnol.* 38:105-140 (1993)). The recursive sequence recombination methods described above can be used to optimize the ability of native hosts or heterologous hosts to utilize a substrate of interest, to evolve more efficient transport systems, to increase or alter specificity for certain substrates, and so on.

4.7.8.10 MODIFICATION OF CELL PROPERTIES

Although not strictly examples of manipulation of intermediary metabolism, recursive sequence recombination techniques can be used to improve or alter other aspects of cell properties, from growth rate to ability to secrete certain desired compounds to ability to tolerate increased temperature or other environmental stresses. Some examples of traits engineered by traditional methods include expression of heterologous proteins in bacteria, yeast, and other eukaryotic cells, antibiotic resistance, and phage resistance. Any of these traits is advantageously evolved by the recursive sequence recombination techniques of the instant invention. Examples include replacement of one nutrient uptake system (e.g. ammonia in *Methylophilus methylotrophus*) with another that is more energy efficient; expression of haemoglobin to improve growth under conditions of limiting oxygen; redirection of toxic metabolic end products to less toxic compounds; expression of genes conferring tolerance to salt, drought and toxic compounds and resistance to pathogens, antibiotics and bacteriophage, reviewed in Cameron et. al. *Appl Biochem Biotechnol*, 38:105-140 (1993).

The heterologous genes encoding these functions all have the potential for further optimization in their new hosts by existing recursive sequence recombination technology. Since these functions increase cell growth rates under the desired growth conditions, optimization of the genes by evolution simply involves recombining the

DNA recursively and selecting the recombinants that grow faster with limiting oxygen, higher toxic compound concentration, or whatever is the appropriate growth condition for the parameter being improved.

Since these functions increase cell growth rates under the desired growth conditions, optimization of the genes by "evolution" can simply involve "stochastic &/or non-stochastic mutagenesis" the DNA and selecting the recombinants that grow faster with limiting oxygen, higher toxic compound concentration or whatever restrictive condition is being overcome. Cultured mammalian cells also require essential amino acids to be present in the growth medium. This requirement could also be circumvented by expression of heterologous metabolic pathways that synthesize these amino acids (Rees et al., Biotechnology 8:629-633 (1990). Recursive sequence recombination would provide a mechanism for optimizing the expression of these genes in mammalian cells. Once again, a preferred selection would be for cells that can grow in the absence of added amino acids.

Yet another candidate for improvement through the techniques of the invention is symbiotic nitrogen fixation. Genes involved in nodulation (nod, ndv), nitrogen reduction (nif, fix), host range determination (nod, hsp), bacteriocin production (tfx), surface polysaccharide synthesis (exo) and energy utilization (dct, hup) which have been identified (Paau, Biotech. Adv. 9:173-184 (1991)). The main function of recursive sequence recombination in this case is in improving the survival of strains that are already known to be better nitrogen fixers. These strains tend to be less good at competing with strains already present in the environment, even though they are better at nitrogen fixation. Targets for recursive sequence recombination such as nodulation and host range determination genes can be modified and selected for by their ability to grow on the new host.

Similarly any bacteriocin or energy utilization genes that will improve the competitiveness of the strain will also result in greater growth rates. Selection can simply be performed by subjecting the target genes to recursive sequence recombination and forcing the inoculant to compete with wild type nitrogen fixing

process.

4.16.5.3.11 Creation of the Initial Population of Sequences

The initial small population of the specific nucleic acid sequences having mutations may be created by a number of different methods. Mutations may be created by error-prone PCR. Error-prone PCR uses low-fidelity polymerization conditions to introduce a low level of point mutations randomly over a long sequence.

Alternatively, mutations can be introduced into the template polynucleotide by oligonucleotide-directed mutagenesis. In oligonucleotide-directed mutagenesis, a short sequence of the polynucleotide is removed from the polynucleotide using restriction enzyme digestion and is replaced with a synthetic polynucleotide in which various bases have been altered from the original sequence. The polynucleotide sequence can also be altered by chemical mutagenesis. Chemical mutagens include, for example, sodium bisulfite, nitrous acid, hydroxylamine, hydrazine or formic acid. Other agents which are analogues of nucleotide precursors include nitrosoguanidine, 5-bromouracil, 2-aminopurine, or acridine. Generally, these agents are added to the PCR reaction in place of the nucleotide precursor thereby mutating the sequence. Intercalating agents such as proflavine, acriflavine, quinacrine and the like can also be used. Random mutagenesis of the polynucleotide sequence can also be achieved by irradiation with X-rays or ultraviolet light. Generally, plasmid polynucleotides so mutagenized are introduced into *E. coli* and propagated as a pool or library of hybrid plasmids.

Alternatively the small mixed population of specific nucleic acids may be found in nature in that they may consist of different alleles of the same gene or the same gene from different related species (*i.e.*, cognate genes). Alternatively, they may be related DNA sequences found within one species, for example, the immunoglobulin genes.

Once the mixed population of the specific nucleic acid sequences is generated, the polynucleotides can be used directly or inserted into an appropriate cloning vector, using techniques well-known in the art.

4.16.5.3.11.1 The Choice of Vector

The choice of vector depends on the size of the polynucleotide sequence and the host cell to be employed in the methods of this invention. The templates of this invention may be plasmids, phages, cosmids, phagemids, viruses (*e.g.*, retroviruses, parainfluenzavirus, herpesviruses, reoviruses, paramyxoviruses, and the like), or selected portions thereof (*e.g.*, coat protein, spike glycoprotein, capsid protein). For example, cosmids and phagemids are preferred where the specific nucleic acid sequence to be mutated is larger because these vectors are able to stably propagate large polynucleotides.

4.16.5.3.11.2 Clonal Amplification

If the mixed population of the specific nucleic acid sequence is cloned into a vector it can be clonally amplified by inserting each vector into a host cell and allowing the host cell to amplify the vector. This is referred to as clonal amplification because while the absolute number of nucleic acid sequences increases, the number of hybrids does not increase. Utility can be readily determined by screening expressed polypeptides.

4.16.5.3.12 Incorporation of Any Sequence Mixture at Any Specific Position

The DNA shuffling method of this invention can be performed blindly on a pool of unknown sequences. By adding to the reassembly mixture oligonucleotides (with ends that are homologous to the sequences being reassembled) any sequence mixture can be incorporated at any specific position into another sequence mixture. Thus, it is contemplated that mixtures of synthetic oligonucleotides, PCR polynucleotides or even whole genes can be mixed into another sequence library at defined positions. The insertion of one sequence (mixture) is independent from the insertion of a sequence in another part of the template. Thus, the degree of recombination, the homology required, and the diversity of the library can be independently and simultaneously varied along the length of the reassembled DNA.

This approach of mixing two genes may be useful for the humanization of antibodies

from murine hybridomas. The approach of mixing two genes or inserting alternative sequences into genes may be useful for any therapeutically used protein, for example, interleukin I, antibodies, tPA and growth hormone. The approach may also be useful in any nucleic acid for example, promoters or introns or 3' untranslated region or 5' untranslated regions of genes to increase expression or alter specificity of expression of proteins. The approach may also be used to mutate ribozymes or aptamers.

4.16.5.3.13 Creation of Scaffold-like Proteins

Shuffling requires the presence of homologous regions separating regions of diversity. Scaffold-like protein structures may be particularly suitable for shuffling. The conserved scaffold determines the overall folding by self-association, while displaying relatively unrestricted loops that mediate the specific binding. Examples of such scaffolds are the immunoglobulin beta-barrel, and the four-helix bundle which are well-known in the art. This shuffling can be used to create scaffold-like proteins with various combinations of mutated sequences for binding.

4.16.5.4 *In vitro* Shuffling

The equivalents of some standard genetic matings may also be performed by shuffling *in vitro*. For example, a "molecular backcross" can be performed by repeatedly mixing the hybrid's nucleic acid with the wild-type nucleic acid while selecting for the mutations of interest. As in traditional breeding, this approach can be used to combine phenotypes from different sources into a background of choice. It is useful, for example, for the removal of neutral mutations that affect unselected characteristics (*i.e.* immunogenicity). Thus it can be useful to determine which mutations in a protein are involved in the enhanced biological activity and which are not, an advantage which cannot be achieved by error-prone mutagenesis or cassette mutagenesis methods.

Large, functional genes can be assembled correctly from a mixture of small random polynucleotides. This reaction may be of use for the reassembly of genes from the

highly fragmented DNA of fossils. In addition random nucleic acid fragments from fossils may be combined with polynucleotides from similar genes from related species.

4.16.5.4.1 *In Vitro* Amplification of a Genome

It is also contemplated that the method of this invention can be used for the *in vitro* amplification of a whole genome from a single cell as is needed for a variety of research and diagnostic applications. DNA amplification by PCR is in practice limited to a length of about 40 kb. Amplification of a whole genome such as that of *E. coli* (5, 000 kb) by PCR would require about 250 primers yielding 125 forty kb polynucleotides. This approach is not practical due to the unavailability of sufficient sequence data. On the other hand, random production of polynucleotides of the genome with sexual PCR cycles, followed by gel purification of small polynucleotides will provide a multitude of possible primers. Use of this mix of random small polynucleotides as primers in a PCR reaction alone or with the whole genome as the template should result in an inverse chain reaction with the theoretical endpoint of a single concatemer containing many copies of the genome.

100 fold amplification in the copy number and an average polynucleotide size of greater than 50 kb may be obtained when only random polynucleotides are used. It is thought that the larger concatemer is generated by overlap of many smaller polynucleotides. The quality of specific PCR products obtained using synthetic primers will be indistinguishable from the product obtained from unamplified DNA. It is expected that this approach will be useful for the mapping of genomes.

The polynucleotide to be shuffled can be produced as random or non-random polynucleotides, at the discretion of the practitioner.

4.16.5.5 *In vivo* Shuffling

In an embodiment of *in vivo* shuffling, the mixed population of the specific nucleic acid sequence is introduced into bacterial or eukaryotic cells under conditions such that at least two different nucleic acid sequences are present in each host cell. The

polynucleotides can be introduced into the host cells by a variety of different methods. The host cells can be transformed with the smaller polynucleotides using methods known in the art, for example treatment with calcium chloride. If the polynucleotides are inserted into a phage genome, the host cell can be transfected with the recombinant phage genome having the specific nucleic acid sequences. Alternatively, the nucleic acid sequences can be introduced into the host cell using electroporation, transfection, lipofection, biolistics, conjugation, and the like.

In general, in this embodiment, the specific nucleic acids sequences will be present in vectors which are capable of stably replicating the sequence in the host cell. In addition, it is contemplated that the vectors will encode a marker gene such that host cells having the vector can be selected. This ensures that the mutated specific nucleic acid sequence can be recovered after introduction into the host cell. However, it is contemplated that the entire mixed population of the specific nucleic acid sequences need not be present on a vector sequence. Rather only a sufficient number of sequences need be cloned into vectors to ensure that after introduction of the polynucleotides into the host cells each host cell contains one vector having at least one specific nucleic acid sequence present therein. It is also contemplated that rather than having a subset of the population of the specific nucleic acids sequences cloned into vectors, this subset may be already stably integrated into the host cell.

4.16.5.5.1 Homologous Recombination

It has been found that when two polynucleotides which have regions of identity are inserted into the host cells homologous recombination occurs between the two polynucleotides. Such recombination between the two mutated specific nucleic acid sequences will result in the production of double or triple hybrids in some situations.

4.16.5.5.2 Increase in the Frequency of Recombination

It has also been found that the frequency of recombination is increased if some of the mutated specific nucleic acid sequences are present on linear nucleic acid molecules.

Therefore, in a preferred embodiment, some of the specific nucleic acid sequences are present on linear polynucleotides.

4.16.5.5.3 Identification of Host Cell Transformants Containing Desired Sequences

After transformation, the host cell transformants are placed under selection to identify those host cell transformants which contain mutated specific nucleic acid sequences having the qualities desired. For example, if increased resistance to a particular drug is desired then the transformed host cells may be subjected to increased concentrations of the particular drug and those transformants producing mutated proteins able to confer increased drug resistance will be selected. If the enhanced ability of a particular protein to bind to a receptor is desired, then expression of the protein can be induced from the transformants and the resulting protein assayed in a ligand binding assay by methods known in the art to identify that subset of the mutated population which shows enhanced binding to the ligand. Alternatively, the protein can be expressed in another system to ensure proper processing.

Once a subset of the first recombined specific nucleic acid sequences (daughter sequences) having the desired characteristics are identified, they are then subject to a second round of recombination.

4.16.5.5.4 The Second Cycle of Recombination

In the second cycle of recombination, the recombined specific nucleic acid sequences may be mixed with the original mutated specific nucleic acid sequences (parent sequences) and the cycle repeated as described above. In this way a set of second recombined specific nucleic acids sequences can be identified which have enhanced characteristics or encode for proteins having enhanced properties. This cycle can be repeated a number of times as desired.

It is also contemplated that in the second or subsequent recombination cycle, a backcross can be performed. A molecular backcross can be performed by mixing the desired specific nucleic acid sequences with a large number of the wild-type sequence, such that at least one wild-type nucleic acid sequence and a mutated nucleic

acid sequence are present in the same host cell after transformation. Recombination with the wild-type specific nucleic acid sequence will eliminate those neutral mutations that may affect unselected characteristics such as immunogenicity but not the selected characteristics.

4.16.5.5.5 Generation of a Subset of the Specific Nucleic Acid Sequences

In another embodiment of this invention, it is contemplated that during the first round a subset of the specific nucleic acid sequences can be generated as smaller polynucleotides by slowing or halting their PCR amplification prior to introduction into the host cell. The size of the polynucleotides must be large enough to contain some regions of identity with the other sequences so as to homologously recombine with the other sequences. The size of the polynucleotides will range from 0.03 kb to 100 kb more preferably from 0.2 kb to 10 kb. It is also contemplated that in subsequent rounds, all of the specific nucleic acid sequences other than the sequences selected from the previous round may be utilized to generate PCR polynucleotides prior to introduction into the host cells.

The shorter polynucleotide sequences can be single-stranded or double-stranded. If the sequences were originally single-stranded and have become double-stranded they can be denatured with heat, chemicals or enzymes prior to insertion into the host cell. The reaction conditions suitable for separating the strands of nucleic acid are well known in the art.

The steps of this process can be repeated indefinitely, being limited only by the number of possible hybrids which can be achieved. After a certain number of cycles, all possible hybrids will have been achieved and further cycles are redundant.

In an embodiment the same mutated template nucleic acid is repeatedly recombined and the resulting recombinants selected for the desired characteristic.

4.16.5.5.6 Cloning into a Vector Capable of Replicating in a Bacteria

Therefore, the initial pool or population of mutated template nucleic acid is cloned into a vector capable of replicating in a bacteria such as *E. coli*. The particular vector

is not essential, so long as it is capable of autonomous replication in *E. coli*. In a preferred embodiment, the vector is designed to allow the expression and production of any protein encoded by the mutated specific nucleic acid linked to the vector. It is also preferred that the vector contain a gene encoding for a selectable marker.

The population of vectors containing the pool of mutated nucleic acid sequences is introduced into the *E. coli* host cells. The vector nucleic acid sequences may be introduced by transformation, transfection or infection in the case of phage. The concentration of vectors used to transform the bacteria is such that a number of vectors is introduced into each cell. Once present in the cell, the efficiency of homologous recombination is such that homologous recombination occurs between the various vectors. This results in the generation of hybrids (daughters) having a combination of mutations which differ from the original parent mutated sequences.

The host cells are then clonally replicated and selected for the marker gene present on the vector. Only those cells having a plasmid will grow under the selection.

4.16.5.5.7 Testing for the Presence of Favorable Mutations

The host cells which contain a vector are then tested for the presence of favorable mutations. Such testing may consist of placing the cells under selective pressure, for example, if the gene to be selected is an improved drug resistance gene. If the vector allows expression of the protein encoded by the mutated nucleic acid sequence, then such selection may include allowing expression of the protein so encoded, isolation of the protein and testing of the protein to determine whether, for example, it binds with increased efficiency to the ligand of interest.

4.16.5.5.8 Isolation of the Desired Nucleic Acid Sequence

Once a particular daughter mutated nucleic acid sequence has been identified which confers the desired characteristics, the nucleic acid is isolated either already linked to the vector or separated from the vector. This nucleic acid is then mixed with the first or parent population of nucleic acids and the cycle is repeated.

It has been shown that by this method nucleic acid sequences having enhanced

desired properties can be selected.

4.16.5.5.9 Addition of Parental Mutated Sequences to the Cells Containing the First Generation of Hybrids

In an alternate embodiment, the first generation of hybrids are retained in the cells and the parental mutated sequences are added again to the cells. Accordingly, the first cycle of Embodiment I is conducted as described above. However, after the daughter nucleic acid sequences are identified, the host cells containing these sequences are retained.

The parent mutated specific nucleic acid population, either as polynucleotides or cloned into the same vector is introduced into the host cells already containing the daughter nucleic acids. Recombination is allowed to occur in the cells and the next generation of recombinants, or granddaughters are selected by the methods described above.

This cycle can be repeated a number of times until the nucleic acid or peptide having the desired characteristics is obtained. It is contemplated that in subsequent cycles, the population of mutated sequences which are added to the preferred hybrids may come from the parental hybrids or any subsequent generation.

4.16.5.5.10 "Molecular" Backcross to Eliminate Any Neutral Mutations

In an alternative embodiment, the invention provides a method of conducting a "molecular" backcross of the obtained recombinant specific nucleic acid in order to eliminate any neutral mutations. Neutral mutations are those mutations which do not confer onto the nucleic acid or peptide the desired properties. Such mutations may however confer on the nucleic acid or peptide undesirable characteristics.

Accordingly, it is desirable to eliminate such neutral mutations. The method of this invention provide a means of doing so.

In this embodiment, after the hybrid nucleic acid, having the desired characteristics, is obtained by the methods of the embodiments, the nucleic acid, the vector having the nucleic acid or the host cell containing the vector and nucleic acid is isolated.

The nucleic acid or vector is then introduced into the host cell with a large excess of

the wild-type nucleic acid. The nucleic acid of the hybrid and the nucleic acid of the wild-type sequence are allowed to recombine. The resulting recombinants are placed under the same selection as the hybrid nucleic acid. Only those recombinants which retained the desired characteristics will be selected. Any silent mutations which do not provide the desired characteristics will be lost through recombination with the wild-type DNA. This cycle can be repeated a number of times until all of the silent mutations are eliminated.

Thus the methods of this invention can be used in a molecular backcross to eliminate unnecessary or silent mutations.

4.16.5.6 Utility

The *in vivo* recombination method of this invention can be performed blindly on a pool of unknown hybrids or alleles of a specific polynucleotide or sequence. However, it is not necessary to know the actual DNA or RNA sequence of the specific polynucleotide.

The approach of using recombination within a mixed population of genes can be useful for the generation of any useful proteins, for example, interleukin I, antibodies, tPA and growth hormone. This approach may be used to generate proteins having altered specificity or activity. The approach may also be useful for the generation of hybrid nucleic acid sequences, for example, promoter regions, introns, exons, enhancer sequences, 3' untranslated regions or 5' untranslated regions of genes. Thus this approach may be used to generate genes having increased rates of expression. This approach may also be useful in the study of repetitive DNA sequences. Finally, this approach may be useful to mutate ribozymes or aptamers.

Scaffold-like regions separating regions of diversity in proteins may be particularly suitable for the methods of this invention. The conserved scaffold determines the overall folding by self-association, while displaying relatively unrestricted loops that mediate the specific binding. Examples of such scaffolds are the immunoglobulin beta barrel, and the four-helix bundle. The methods of this invention can be used to

create scaffold-like proteins with various combinations of mutated sequences for binding.

The equivalents of some standard genetic matings may also be performed by the methods of this invention. For example, a "molecular" backcross can be performed by repeated mixing of the hybrid's nucleic acid with the wild-type nucleic acid while selecting for the mutations of interest. As in traditional breeding, this approach can be used to combine phenotypes from different sources into a background of choice. It is useful, for example, for the removal of neutral mutations that affect unselected characteristics (*i.e.* immunogenicity). Thus it can be useful to determine which mutations in a protein are involved in the enhanced biological activity and which are not.

4.16.5.7 Peptide Display Methods

The present method can be used to shuffle, by *in vitro* and/or *in vivo* recombination by any of the disclosed methods, and in any combination, polynucleotide sequences selected by peptide display methods, wherein an associated polynucleotide encodes a displayed peptide which is screened for a phenotype (*e.g.*, for affinity for a predetermined receptor (ligand).

An increasingly important aspect of bio-pharmaceutical drug development and molecular biology is the identification of peptide structures, including the primary amino acid sequences, of peptides or peptidomimetics that interact with biological macromolecules. One method of identifying peptides that possess a desired structure or functional property, such as binding to a predetermined biological macromolecule (*e.g.*, a receptor), involves the screening of a large library of peptides for individual library members which possess the desired structure or functional property conferred by the amino acid sequence of the peptide.

In addition to direct chemical synthesis methods for generating peptide libraries, several recombinant DNA methods also have been reported. One type involves the display of a peptide sequence, antibody, or other protein on the surface of a

bacteriophage particle or cell. Generally, in these methods each bacteriophage particle or cell serves as an individual library member displaying a single species of displayed peptide in addition to the natural bacteriophage or cell protein sequences. Each bacteriophage or cell contains the nucleotide sequence information encoding the particular displayed peptide sequence; thus, the displayed peptide sequence can be ascertained by nucleotide sequence determination of an isolated library member.

A well-known peptide display method involves the presentation of a peptide sequence on the surface of a filamentous bacteriophage, typically as a fusion with a bacteriophage coat protein. The bacteriophage library can be incubated with an immobilized, predetermined macromolecule or small molecule (e.g., a receptor) so that bacteriophage particles which present a peptide sequence that binds to the immobilized macromolecule can be differentially partitioned from those that do not present peptide sequences that bind to the predetermined macromolecule. The bacteriophage particles (*i.e.*, library members) which are bound to the immobilized macromolecule are then recovered and replicated to amplify the selected bacteriophage sub-population for a subsequent round of affinity enrichment and phage replication. After several rounds of affinity enrichment and phage replication, the bacteriophage library members that are thus selected are isolated and the nucleotide sequence encoding the displayed peptide sequence is determined, thereby identifying the sequence(s) of peptides that bind to the predetermined macromolecule (e.g., receptor). Such methods are further described in PCT patent publication Nos. 91/17271, 91/18980, and 91/19818 and 93/08278.

The latter PCT publication describes a recombinant DNA method for the display of peptide ligands that involves the production of a library of fusion proteins with each fusion protein composed of a first polypeptide portion, typically comprising a variable sequence, that is available for potential binding to a predetermined macromolecule, and a second polypeptide portion that binds to DNA, such as the DNA vector encoding the individual fusion protein. When transformed host cells are cultured under conditions that allow for expression of the fusion protein, the fusion protein binds to the DNA vector encoding it. Upon lysis of the host cell, the fusion protein/vector DNA complexes can be screened against a predetermined

macromolecule in much the same way as bacteriophage particles are screened in the phage-based display system, with the replication and sequencing of the DNA vectors in the selected fusion protein/vector DNA complexes serving as the basis for identification of the selected library peptide sequence(s).

4.16.5.7.1 Hybrid Methods for Generating Libraries of Peptides and Like Polymers

Other systems for generating libraries of peptides and like polymers have aspects of both the recombinant and *in vitro* chemical synthesis methods. In these hybrid methods, cell-free enzymatic machinery is employed to accomplish the *in vitro* synthesis of the library members (*i.e.*, peptides or polynucleotides). In one type of method, RNA molecules with the ability to bind a predetermined protein or a predetermined dye molecule were selected by alternate rounds of selection and PCR amplification (Tuerk and Gold (1990) *Science* 249: 505; Ellington and Szostak (1990) *Nature* 346: 818). A similar technique was used to identify DNA sequences which bind a predetermined human transcription factor (Thiesen and Bach (1990) *Nucleic Acids Res.* 18: 3203; Beaudry and Joyce (1992) *Science* 257: 635; PCT patent publication Nos. 92/05258 and 92/14843). In a similar fashion, the technique of *in vitro* translation has been used to synthesize proteins of interest and has been proposed as a method for generating large libraries of peptides. These methods which rely upon *in vitro* translation, generally comprising stabilized polysome complexes, are described further in PCT patent publication Nos. 88/08453, 90/05785, 90/07003, 91/02076, 91/05058, and 92/02536. Applicants have described methods in which library members comprise a fusion protein having a first polypeptide portion with DNA binding activity and a second polypeptide portion having the library member unique peptide sequence; such methods are suitable for use in cell-free *in vitro* selection formats, among others.

4.16.5.7.2 The Displayed Peptide Sequences

The displayed peptide sequences can be of varying lengths, typically from 3-5000 amino acids long or longer, frequently from 5-100 amino acids long, and often from

about 8-15 amino acids long. A library can comprise library members having varying lengths of displayed peptide sequence, or may comprise library members having a fixed length of displayed peptide sequence. Portions or all of the displayed peptide sequence(s) can be random, pseudorandom, defined set kernel, fixed, or the like. The present display methods include methods for *in vitro* and *in vivo* display of single-chain antibodies, such as nascent scFv on polysomes or scfv displayed on phage, which enable large-scale screening of scfv libraries having broad diversity of variable region sequences and binding specificities.

4.16.5.7.3 Sequence Framework Peptide Libraries

The present invention also provides random, pseudorandom, and defined sequence framework peptide libraries and methods for generating and screening those libraries to identify useful compounds (*e.g.*, peptides, including single-chain antibodies) that bind to receptor molecules or epitopes of interest or gene products that modify peptides or RNA in a desired fashion. The random, pseudorandom, and defined sequence framework peptides are produced from libraries of peptide library members that comprise displayed peptides or displayed single-chain antibodies attached to a polynucleotide template from which the displayed peptide was synthesized. The mode of attachment may vary according to the specific embodiment of the invention selected, and can include encapsulation in a phage particle or incorporation in a cell.

4.16.5.7.4 Selecting for the Desired Peptide Using Affinity Enrichment

A method of affinity enrichment allows a very large library of peptides and single-chain antibodies to be screened and the polynucleotide sequence encoding the desired peptide(s) or single-chain antibodies to be selected. The polynucleotide can then be isolated and shuffled to recombine combinatorially the amino acid sequence of the selected peptide(s) (or predetermined portions thereof) or single-chain antibodies (or just VHI, VLI or CDR portions thereof). Using these methods, one can identify a peptide or single-chain antibody as having a desired binding affinity for a molecule and can exploit the process of shuffling to converge rapidly to a desired high-affinity peptide or scfv. The peptide or antibody can then be synthesized in bulk

by conventional means for any suitable use (*e.g.*, as a therapeutic or diagnostic agent).

A significant advantage of the present invention is that no prior information regarding an expected ligand structure is required to isolate peptide ligands or antibodies of interest. The peptide identified can have biological activity, which is meant to include at least specific binding affinity for a selected receptor molecule and, in some instances, will further include the ability to block the binding of other compounds, to stimulate or inhibit metabolic pathways, to act as a signal or messenger, to stimulate or inhibit cellular activity, and the like.

4.16.5.7.5 Shuffling Sequences Selected by Affinity Screening

The present invention also provides a method for shuffling a pool of polynucleotide sequences selected by affinity screening a library of polysomes displaying nascent peptides (including single-chain antibodies) for library members which bind to a predetermined receptor (*e.g.*, a mammalian proteinaceous receptor such as, for example, a peptidergic hormone receptor, a cell surface receptor, an intracellular protein which binds to other protein(s) to form intracellular protein complexes such as hetero-dimers and the like) or epitope (*e.g.*, an immobilized protein, glycoprotein, oligosaccharide, and the like).

Polynucleotide sequences selected in a first selection round (typically by affinity selection for binding to a receptor (*e.g.*, a ligand)) by any of these methods are pooled and the pool(s) is/are shuffled by *in vitro* and/or *in vivo* recombination to produce a shuffled pool comprising a population of recombinant selected polynucleotide sequences. The recombinant selected polynucleotide sequences are subjected to at least one subsequent selection round. The polynucleotide sequences selected in the subsequent selection round(s) can be used directly, sequenced, and/or subjected to one or more additional rounds of shuffling and subsequent selection. Selected sequences can also be back-crossed with polynucleotide sequences encoding neutral sequences (*i.e.*, having insubstantial functional effect on binding), such as for example by back-crossing with a wild-type or naturally-occurring sequence substantially identical to a selected sequence to produce native-like functional peptides, which may be less

immunogenic. Generally, during back-crossing subsequent selection is applied to retain the property of binding to the predetermined receptor (ligand).

Prior to or concomitant with the shuffling of selected sequences, the sequences can be mutagenized. In one embodiment, selected library members are cloned in a prokaryotic vector (*e.g.*, plasmid, phagemid, or bacteriophage) wherein a collection of individual colonies (or plaques) representing discrete library members are produced. Individual selected library members can then be manipulated (*e.g.*, by site-directed mutagenesis, cassette mutagenesis, chemical mutagenesis, PCR mutagenesis, and the like) to generate a collection of library members representing a kernal of sequence diversity based on the sequence of the selected library member. The sequence of an individual selected library member or pool can be manipulated to incorporate random mutation, pseudorandom mutation, defined kernal mutation (*i.e.*, comprising variant and invariant residue positions and/or comprising variant residue positions which can comprise a residue selected from a defined subset of amino acid residues), codon-based mutation, and the like, either segmentally or over the entire length of the individual selected library member sequence. The mutagenized selected library members are then shuffled by *in vitro* and/or *in vivo* recombinatorial shuffling as disclosed herein.

4.16.5.7.6 Peptide Libraries Comprising a Plurality of Individual Library Members

The invention also provides peptide libraries comprising a plurality of individual library members of the invention, wherein (1) each individual library member of said plurality comprises a sequence produced by shuffling of a pool of selected sequences, and (2) each individual library member comprises a variable peptide segment sequence or single-chain antibody segment sequence which is distinct from the variable peptide segment sequences or single-chain antibody sequences of other individual library members in said plurality (although some library members may be present in more than one copy per library due to uneven amplification, stochastic probability, or the like).

4.16.5.7.7 Product-by-Process

The invention also provides a product-by-process, wherein selected polynucleotide sequences having (or encoding a peptide having) a predetermined binding specificity are formed by the process of: (1) screening a displayed peptide or displayed single-chain antibody library against a predetermined receptor (*e.g.*, ligand) or epitope (*e.g.*, antigen macromolecule) and identifying and/or enriching library members which bind to the predetermined receptor or epitope to produce a pool of selected library members, (2) shuffling by recombination the selected library members (or amplified or cloned copies thereof) which binds the predetermined epitope and has been thereby isolated and/or enriched from the library to generate a shuffled library, and (3) screening the shuffled library against the predetermined receptor (*e.g.*, ligand) or epitope (*e.g.*, antigen macromolecule) and identifying and/or enriching shuffled library members which bind to the predetermined receptor or epitope to produce a pool of selected shuffled library members.

4.16.5.8 Antibody Display and Screening Methods

The present method can be used to shuffle, by *in vitro* and/or *in vivo* recombination by any of the disclosed methods, and in any combination, polynucleotide sequences selected by antibody display methods, wherein an associated polynucleotide encodes a displayed antibody which is screened for a phenotype (*e.g.*, for affinity for binding a predetermined antigen (ligand)).

Various molecular genetic approaches have been devised to capture the vast immunological repertoire represented by the extremely large number of distinct variable regions which can be present in immunoglobulin chains. The naturally-occurring germ line immunoglobulin heavy chain locus is composed of separate tandem arrays of variable segment genes located upstream of a tandem array of diversity segment genes, which are themselves located upstream of a tandem array of joining (i) region genes, which are located upstream of the constant region genes. During B lymphocyte development, V-D-J rearrangement occurs wherein a heavy chain variable region gene (VH) is formed by rearrangement to form a fused D

segment followed by rearrangement with a V segment to form a V-D-J joined product gene which, if productively rearranged, encodes a functional variable region (VH) of a heavy chain. Similarly, light chain loci rearrange one of several V segments with one of several J segments to form a gene encoding the variable region (VL) of a light chain.

4.16.5.8.1 Sequence Diversity

The vast repertoire of variable regions possible in immunoglobulins derives in part from the numerous combinatorial possibilities of joining V and i segments (and, in the case of heavy chain loci, D segments) during rearrangement in B cell development. Additional sequence diversity in the heavy chain variable regions arises from non-uniform rearrangements of the D segments during V-D-J joining and from N region addition. Further, antigen-selection of specific B cell clones selects for higher affinity variants having non-germline mutations in one or both of the heavy and light chain variable regions; a phenomenon referred to as "affinity maturation" or "affinity sharpening". Typically, these "affinity sharpening" mutations cluster in specific areas of the variable region, most commonly in the complementarity-determining regions (CDRs).

4.16.5.8.2 Prokaryotic Expression Systems

In order to overcome many of the limitations in producing and identifying high-affinity immunoglobulins through antigen-stimulated B cell development (*i.e.*, immunization), various prokaryotic expression systems have been developed that can be manipulated to produce combinatorial antibody libraries which may be screened for high-affinity antibodies to specific antigens. Recent advances in the expression of antibodies in *Escherichia coli* and bacteriophage systems (*see*, "Alternative Peptide Display Methods", *infra*) have raised the possibility that virtually any specificity can be obtained by either cloning antibody genes from characterized hybridomas or by *de novo* selection using antibody gene libraries (*e.g.*, from Ig CDNA).

Combinatorial libraries of antibodies have been generated in bacteriophage lambda expression systems which may be screened as bacteriophage plaques or as colonies of lysogens (Huse *et al.* (1989) Science 246: 1275; Caton and Koprowski (1990) Proc. Natl. Acad. Sci. (U.S.A.) 87: 6450; Mullinax et al (1990) Proc. Natl. Acad. Sci. (U.S.A.) 87: 8095; Persson *et al.* (1991) Proc. Natl. Acad. Sci. (T.J.S.A.) 88: 2432). Various embodiments of bacteriophage antibody display libraries and lambda phage expression libraries have been described (Kang *et al.* (1991) Proc. Natl. Acad. Sci. (U.S.A.) 88: 4363; Clackson *et al.* (1991) Nature 352: 624; McCafferty *et al.* (1990) Nature 348: 552; Burton *et al.* (1991) Proc. Natl. Acad. Sci. (U.S.A.) 88: 10134; Hoogenboom *et al.* (1991) Nucleic Acids Res. 19: 4133; Chang *et al.* (1991) J. Immunol. 147: 3610; Breitling *et al.* (1991) Gene 104: 147; Marks *et al.* (1991) J. Mol. Biol. 222@: 581; Barbas *et al.* (1992) Proc. Natl. Acad. Sci. (U.S.A.) 89: 4457; Hawkins and Winter (1992) J. Immunol. 22: 867; Marks *et al.* (1992) Biotechnology 10: 779; Marks *et al.* (1992) J. Biol. Chem. 267: 16007; Lowman et al (1991) Biochemistry 30: 10832; Lerner *et al.* (1992) Science 258: 1313, incorporated herein by reference). Typically, a bacteriophage antibody display library is screened with a receptor (*e.g.*, polypeptide, carbohydrate, glycoprotein, nucleic acid) that is immobilized (*e.g.*, by covalent linkage to a chromatography resin to enrich for reactive phage by affinity chromatography) and/or labeled (*e.g.*, to screen plaque or colony lifts).

4.16.5.8.3 Single-Chain Fragment Variable Libraries

One particularly advantageous approach has been the use of so-called single-chain fragment variable (scfv) libraries (Marks *et al.* (1992) Biotechnology 10: 779; Winter G and Milstein C (1991) Nature 349: 293; Clackson *et al.* (1991) op. cit.; Marks *et al.* (1991) J. Mol. Biol. 222: 581; Chaudhary *et al.* (1990) Proc. Natl. Acad. Sci. (USA) 87: 1066; Chiswell *et al.* (1992) TIBTECH 10: 80; McCafferty *et al.* (1990) op.cit.; and Huston et al- (1988) Proc. Natl. Acad. Sci. (USA) 85: 5879). Various embodiments of scfv libraries displayed on bacteriophage coat proteins have been described.

Beginning in 1988, single-chain analogues of Fv fragments and their fusion proteins have been reliably generated by antibody engineering methods. The first step generally involves obtaining the genes encoding VH and VL domains with desired binding properties; these V genes may be isolated from a specific hybridoma cell line, selected from a combinatorial V-gene library, or made by V gene synthesis. The single-chain Fv is formed by connecting the component V genes with an oligonucleotide that encodes an appropriately designed linker peptide, such as (Gly-Gly-Gly-Gly-Ser)3 or equivalent linker peptide(s). The linker bridges the C-terminus of the first V region and N-terminus of the second, ordered as either VH-linker-VL or VL-linker-VH'. In principle, the scfv binding site can faithfully replicate both the affinity and specificity of its parent antibody combining site.

Thus, scfv fragments are comprised of VH and VL domains linked into a single polypeptide chain by a flexible linker peptide. After the scfv genes are assembled, they are cloned into a phagemid and expressed at the tip of the M13 phage (or similar filamentous bacteriophage) as fusion proteins with the bacteriophage PIII (gene 3) coat protein. Enriching for phage expressing an antibody of interest is accomplished by panning the recombinant phage displaying a population scfv for binding to a predetermined epitope (e.g., target antigen, receptor).

The linked polynucleotide of a library member provides the basis for replication of the library member after a screening or selection procedure, and also provides the basis for the determination, by nucleotide sequencing, of the identity of the displayed peptide sequence or VH and VL amino acid sequence. The displayed peptide (s) or single-chain antibody (e.g., scfv) and/or its VH and VL domains or their CDRs can be cloned and expressed in a suitable expression system. Often polynucleotides encoding the isolated VH and VL domains will be ligated to polynucleotides encoding constant regions (CH and CL) to form polynucleotides encoding complete antibodies (e.g., chimeric or fully-human), antibody fragments, and the like. Often polynucleotides encoding the isolated CDRs will be grafted into polynucleotides encoding a suitable variable region framework (and optionally constant regions) to form polynucleotides encoding complete antibodies (e.g., humanized or fully-human), antibody fragments, and the like. Antibodies can be used to isolate preparative

quantities of the antigen by immunoaffinity chromatography. Various other uses of such antibodies are to diagnose and/or stage disease (*e.g.*, neoplasia) and for therapeutic application to treat disease, such as for example: neoplasia, autoimmune disease, AIDS, cardiovascular disease, infections, and the like.

4.16.5.8.4 Increasing the Combinatorial Diversity of a SCFV Library

Various methods have been reported for increasing the combinatorial diversity of a scfv library to broaden the repertoire of binding species (idiotype spectrum). The use of PCR has permitted the variable regions to be rapidly cloned either from a specific hybridoma source or as a gene library from non-immunized cells, affording combinatorial diversity in the assortment of VH and VL cassettes which can be combined. Furthermore, the VH and VL cassettes can themselves be diversified, such as by random, pseudorandom, or directed mutagenesis. Typically, VH and VL cassettes are diversified in or near the complementarity-determining regions (CDRS), often the third CDR, CDR3. Enzymatic inverse PCR mutagenesis has been shown to be a simple and reliable method for constructing relatively large libraries of scfv site-directed hybrids (Stemmer *et al.* (1993) *Biotechniques* 14: 256), as has error-prone PCR and chemical mutagenesis (Deng *et al.* (1994) *J. Biol. Chem.* 269: 953 3). Riechmann *et al.* (1993) *Biochemistry* 32: 8848 showed semi-rational design of an antibody scfv fragment using site-directed randomization by degenerate oligonucleotide PCR and subsequent phage display of the resultant scfv hybrids. Barbas *et al.* (1992) *on.cit.* attempted to circumvent the problem of limited repertoire sizes resulting from using biased variable region sequences by randomizing the sequence in a synthetic CDR region of a human tetanus toxoid-binding Fab.

CDR randomization has the potential to create approximately 1×10^{20} CDRs for the heavy chain CDR3 alone, and a roughly similar number of variants of the heavy chain CDR1 and CDR2, and light chain CDR1-3 variants. Taken individually or together, the combination possibilities of CDR randomization of heavy and/or light chains requires generating a prohibitive number of bacteriophage clones to produce a clone library representing all possible combinations, the vast majority of which will be non-

binding. Generation of such large numbers of primary transformants is not feasible with current transformation technology and bacteriophage display systems. For example, Barbas *et al.* (1992) *op.cit.* only generated 5×10^7 transformants, which represents only a tiny fraction of the potential diversity of a library of thoroughly randomized CDRS.

Despite these substantial limitations, bacteriophage display of scfv have already yielded a variety of useful antibodies and antibody fusion proteins. A bispecific single chain antibody has been shown to mediate efficient tumor cell lysis (Gruber *et al.* (1994) *J. Immunol.* 152: 5368). Intracellular expression of an anti-Rev scfv has been shown to inhibit HIV-1 virus replication *in vitro* (Duan *et al.* (1994) *Proc. Natl. Acad. Sci. (USA)* 91: 5075), and intracellular expression of an anti-p21rar, scfv has been shown to inhibit meiotic maturation of *Xenopus* oocytes (Biocca *et al.* (1993) *Biochem. Biosphys. Res. Commun.* 197: 422. Recombinant scfv which can be used to diagnose HIV infection have also been reported, demonstrating the diagnostic utility of scfv (Lilley *et al.* (1994) *J. Immunol. Meth.* 171: 211). Fusion proteins wherein an scFv is linked to a second polypeptide, such as a toxin or fibrinolytic activator protein, have also been reported (Holvost *et al.* (1992) *Eur. J. Biochess.* 210: 945; Nicholls *et al.* (1993) *J. Biol. Chem.* 268: 5302).

4.16.5.8.5 Use of *in vitro* and *in vivo* Shuffling Methods to Recombine CDRs

If it were possible to generate scfv libraries having broader antibody diversity and overcoming many of the limitations of conventional CDR mutagenesis and randomization methods which can cover only a very tiny fraction of the potential sequence combinations, the number and quality of scfv antibodies suitable for therapeutic and diagnostic use could be vastly improved. To address this, the *in vitro* and *in vivo* shuffling methods of the invention are used to recombine CDRs which have been obtained (typically via PCR amplification or cloning) from nucleic acids obtained from selected displayed antibodies. Such displayed antibodies can be displayed on cells, on bacteriophage particles, on polysomes, or any suitable antibody display system wherein the antibody is associated with its encoding nucleic acid(s).

In a variation, the CDRs are initially obtained from mRNA (or cDNA) from antibody-producing cells (e.g., plasma cells/splenocytes from an immunized wild-type mouse, a human, or a transgenic mouse capable of making a human antibody as in W092/03918, W093/12227, and W094/25585), including hybridomas derived therefrom.

Polynucleotide sequences selected in a first selection round (typically by affinity selection for displayed antibody binding to an antigen (e.g., a ligand) by any of these methods are pooled and the pool(s) is/are shuffled by *in vitro* and/or *in vivo* recombination, especially shuffling of CDRs (typically shuffling heavy chain CDRs with other heavy chain CDRs and light chain CDRs with other light chain CDRS) to produce a shuffled pool comprising a population of recombined selected polynucleotide sequences. The recombined selected polynucleotide sequences are expressed in a selection format as a displayed antibody and subjected to at least one subsequent selection round. The polynucleotide sequences selected in the subsequent selection round(s) can be used directly, sequenced, and/or subjected to one or more additional rounds of shuffling and subsequent selection until an antibody of the desired binding affinity is obtained. Selected sequences can also be back-crossed with polynucleotide sequences encoding neutral antibody framework sequences (*i.e.*, having insubstantial functional effect on antigen binding), such as for example by back-crossing with a human variable region framework to produce human-like sequence antibodies. Generally, during back-crossing subsequent selection is applied to retain the property of binding to the predetermined antigen.

4.16.5.8.6 Controlling the Average Binding Affinity of Selected SCFV Library Members

Alternatively, or in combination with the noted variations, the valency of the target epitope may be varied to control the average binding affinity of selected scfv library members. The target epitope can be bound to a surface or substrate at varying densities, such as by including a competitor epitope, by dilution, or by other method known to those in the art. A high density (valency) of predetermined epitope can be

used to enrich for scfv library members which have relatively low affinity, whereas a low density (valency) can preferentially enrich for higher affinity scfv library members.

4.16.5.8.7 Generating Diverse Variable Segments

For generating diverse variable segments, a collection of synthetic oligonucleotides encoding random, pseudorandom, or a defined sequence kernal set of peptide sequences can be inserted by ligation into a predetermined site (*e.g.*, a CDR). Similarly, the sequence diversity of one or more CDRs of the single-chain antibody cassette(s) can be expanded by mutating the CDR(s) with site-directed mutagenesis, CDR-replacement, and the like. The resultant DNA molecules can be propagated in a host for cloning and amplification prior to shuffling, or can be used directly (*i.e.*, may avoid loss of diversity which may occur upon propagation in a host cell) and the selected library members subsequently shuffled.

Displayed peptide/polynucleotide complexes (library members) which encode a variable segment peptide sequence of interest or a single-chain antibody of interest are selected from the library by an affinity enrichment technique. This is accomplished by means of a immobilized macromolecule or epitope specific for the peptide sequence of interest, such as a receptor, other macromolecule, or other epitope species. Repeating the affinity selection procedure provides an enrichment of library members encoding the desired sequences, which may then be isolated for pooling and shuffling, for sequencing, and/or for further propagation and affinity enrichment.

The library members without the desired specificity are removed by washing. The degree and stringency of washing required will be determined for each peptide sequence or single-chain antibody of interest and the immobilized predetermined macromolecule or epitope. A certain degree of control can be exerted over the binding characteristics of the nascent peptide/DNA complexes recovered by adjusting the conditions of the binding incubation and the subsequent washing. The temperature, pH, ionic strength, divalent cations concentration, and the volume and duration of the washing will select for nascent peptide/DNA complexes within particular ranges of affinity for the immobilized macromolecule. Selection based on

slow dissociation rate, which is usually predictive of high affinity, is often the most practical route. This may be done either by continued incubation in the presence of a saturating amount of free predetermined macromolecule, or by increasing the volume, number, and length of the washes. In each case, the rebinding of dissociated nascent peptide/DNA or peptide/RNA complex is prevented, and with increasing time, nascent peptide/DNA or peptide/RNA complexes of higher and higher affinity are recovered.

Additional modifications of the binding and washing procedures may be applied to find peptides with special characteristics. The affinities of some peptides are dependent on ionic strength or cation concentration. This is a useful characteristic for peptides that will be used in affinity purification of various proteins when gentle conditions for removing the protein from the peptides are required.

One variation involves the use of multiple binding targets (multiple epitope species, multiple receptor species), such that a scfv library can be simultaneously screened for a multiplicity of scfv which have different binding specificities. Given that the size of a scfv library often limits the diversity of potential scfv sequences, it is typically desirable to us scfv libraries of as large a size as possible. The time and economic considerations of generating a number of very large polysome scFv-display libraries can become prohibitive. To avoid this substantial problem, multiple predetermined epitope species (receptor species) can be concomitantly screened in a single library, or sequential screening against a number of epitope species can be used. In one variation, multiple target epitope species, each encoded on a separate bead (or subset of beads), can be mixed and incubated with a polysome-display scfv library under suitable binding conditions. The collection of beads, comprising multiple epitope species, can then be used to isolate, by affinity selection, scfv library members. Generally, subsequent affinity screening rounds can include the same mixture of beads, subsets thereof, or beads containing only one or two individual epitope species. This approach affords efficient screening, and is compatible with laboratory automation, batch processing, and high throughput screening methods.

4.16.5.8.8 Techniques Used to Diversify a Peptide Library or Single-Chain Antibody Library

A variety of techniques can be used in the present invention to diversify a peptide library or single-chain antibody library, or to diversify, prior to or concomitant with shuffling, around variable segment peptides found in early rounds of panning to have sufficient binding activity to the predetermined macromolecule or epitope. In one approach, the positive selected peptide/polynucleotide complexes (those identified in an early round of affinity enrichment) are sequenced to determine the identity of the active peptides. Oligonucleotides are then synthesized based on these active peptide sequences, employing a low level of all bases incorporated at each step to produce slight variations of the primary oligonucleotide sequences. This mixture of (slightly) degenerate oligonucleotides is then cloned into the variable segment sequences at the appropriate locations. This method produces systematic, controlled variations of the starting peptide sequences, which can then be shuffled. It requires, however, that individual positive nascent peptide/polynucleotide complexes be sequenced before mutagenesis, and thus is useful for expanding the diversity of small numbers of recovered complexes and selecting variants having higher binding affinity and/or higher binding specificity. In a variation, mutagenic PCR amplification of positive selected peptide/polynucleotide complexes (especially of the variable region sequences, the amplification products of which are shuffled *in vitro* and/or *in vivo* and one or more additional rounds of screening is done prior to sequencing. The same general approach can be employed with single-chain antibodies in order to expand the diversity and enhance the binding affinity/specification, typically by diversifying CDRs or adjacent framework regions prior to or concomitant with shuffling. If desired, shuffling reactions can be spiked with mutagenic oligonucleotides capable of *in vitro* recombination with the selected library members can be included. Thus, mixtures of synthetic oligonucleotides and PCR produced polynucleotides (synthesized by error-prone or high-fidelity methods) can be added to the *in vitro* shuffling mix and be incorporated into resulting shuffled library members (shufflants).

4.16.5.8.9 Generation of a Library of CDR-Varient Single-Chain Antibodies

The present invention of shuffling enables the generation of a vast library of CDR-variant single-chain antibodies. One way to generate such antibodies is to insert synthetic CDRs into the single-chain antibody and/or CDR randomization prior to or

concomitant with shuffling. The sequences of the synthetic CDR cassettes are selected by referring to known sequence data of human CDR and are selected in the discretion of the practitioner according to the following guidelines: synthetic CDRs will have at least 40 percent positional sequence identity to known CDR sequences, and preferably will have at least 50 to 70 percent positional sequence identity to known CDR sequences. For example, a collection of synthetic CDR sequences can be generated by synthesizing a collection of oligonucleotide sequences on the basis of naturally-occurring human CDR sequences listed in Kabat *et al.* (1991) op. cit.; the pool(s) of synthetic CDR sequences are calculated to encode CDR peptide sequences having at least 40 percent sequence identity to at least one known naturally-occurring human CDR sequence. Alternatively, a collection of naturally-occurring CDR sequences may be compared to generate consensus sequences so that amino acids used at a residue position frequently (*i.e.*, in at least 5 percent of known CDR sequences) are incorporated into the synthetic CDRs at the corresponding position(s). Typically, several (*e.g.*, 3 to about 50) known CDR sequences are compared and observed natural sequence variations between the known CDRs are tabulated, and a collection of oligonucleotides encoding CDR peptide sequences encompassing all or most permutations of the observed natural sequence variations is synthesized. For example but not for limitation, if a collection of human VH CDR sequences have carboxy-terminal amino acids which are either Tyr, Val, Phe, or Asp, then the pool(s) of synthetic CDR oligonucleotide sequences are designed to allow the carboxy-terminal CDR residue to be any of these amino acids. In some embodiments, residues other than those which naturally-occur at a residue position in the collection of CDR sequences are incorporated: conservative amino acid substitutions are frequently incorporated and up to 5 residue positions may be varied to incorporate non-conservative amino acid substitutions as compared to known naturally-occurring CDR sequences. Such CDR sequences can be used in primary library members (prior to first round screening) and/or can be used to spike *in vitro* shuffling reactions of selected library member sequences. Construction of such pools of defined and/or degenerate sequences will be readily accomplished by those of ordinary skill in the art.

The collection of synthetic CDR sequences comprises at least one member that is not known to be a naturally-occurring CDR sequence. It is within the discretion of the practitioner to include or not include a portion of random or pseudorandom sequence corresponding to N region addition in the heavy chain CDR; the N region sequence ranges from 1 nucleotide to about 4 nucleotides occurring at V-D and D-J junctions. A collection of synthetic heavy chain CDR sequences comprises at least about 100 unique CDR sequences, typically at least about 1,000 unique CDR sequences, preferably at least about 10,000 unique CDR sequences, frequently more than 50,000 unique CDR sequences; however, usually not more than about 1×10^6 unique CDR sequences are included in the collection, although occasionally 1×10^7 to 1×10^8 unique CDR sequences are present, especially if conservative amino acid substitutions are permitted at positions where the conservative amino acid substituent is not present or is rare (*i.e.*, less than 0.1 percent) in that position in naturally-occurring human CDRS. In general, the number of unique CDR sequences included in a library should not exceed the expected number of primary transformants in the library by more than a factor of 10. Such single-chain antibodies generally bind of about at least 1×10^{-m} , preferably with an affinity of about at least 5×10^{-7} M-1, more preferably with an affinity of at least 1×10^{-8} M-1 to 1×10^{-9} M-1 or more, sometimes up to 1×10^{-10} M-1 or more. Frequently, the predetermined antigen is a human protein, such as for example a human cell surface antigen (*e.g.*, CD4, CD8, IL-2 receptor, EGF receptor, PDGF receptor), other human biological macromolecule (*e.g.*, thrombomodulin, protein C, carbohydrate antigen, sialyl Lewis antigen, Lselectin), or nonhuman disease associated macromolecule (*e.g.*, bacterial LPS, virion capsid protein or envelope glycoprotein) and the like.

4.16.5.8.10 Expression systems

High affinity single-chain antibodies of the desired specificity can be engineered and expressed in a variety of systems. For example, scfv have been produced in plants (Firek *et al.* (1993) Plant Mol. Biol. 23: 861) and can be readily made in prokaryotic

systems (Owens RJ and Young RJ (1994) *J. Immunol. Meth.* 168: 149; Johnson S and Bird RE (1991) *Methods Enzymol.* 203: 88). Furthermore, the single-chain antibodies can be used as a basis for constructing whole antibodies or various fragments thereof (Kettleborough *et al.* (1994) *Eur. J. Immunol.* 24: 952). The variable region encoding sequence may be isolated (*e.g.*, by PCR amplification or subcloning) and spliced to a sequence encoding a desired human constant region to encode a human sequence antibody more suitable for human therapeutic uses where immunogenicity is preferably minimized. The polynucleotide(s) having the resultant fully human encoding sequence(s) can be expressed in a host cell (*e.g.*, from an expression vector in a mammalian cell) and purified for pharmaceutical formulation.

The DNA expression constructs will typically include an expression control DNA sequence operably linked to the coding sequences, including naturally-associated or heterologous promoter regions. Preferably, the expression control sequences will be eukaryotic promoter systems in vectors capable of transforming or transfecting eukaryotic host cells. Once the vector has been incorporated into the appropriate host, the host is maintained under conditions suitable for high level expression of the nucleotide sequences, and the collection and purification of the mutant' "engineered" antibodies.

As stated previously, the DNA sequences will be expressed in hosts after the sequences have been operably linked to an expression control sequence (*i.e.*, positioned to ensure the transcription and translation of the structural gene). These expression vectors are typically replicable in the host organisms either as episomes or as an integral part of the host chromosomal DNA. Commonly, expression vectors will contain selection markers, *e.g.*, tetracycline or neomycin, to permit detection of those cells transformed with the desired DNA sequences (*see, e.g.*, U.S. Patent 4,704,362, which is incorporated herein by reference).

4.16.5.8.11 Mammalian Tissue Cell Culture

In addition to eukaryotic microorganisms such as yeast, mammalian tissue cell culture

may also be used to produce the polypeptides of the present invention (see, Winnacker, "From Genes to Clones," VCH Publishers, N.Y., N.Y. (1987), which is incorporated herein by reference). Eukaryotic cells are actually preferred, because a number of suitable host cell lines capable of secreting intact immunoglobulins have been developed in the art, and include the CHO cell lines, various COS cell lines, HeLa cells, and myeloma cell lines, but preferably transformed Bcells or hybridomas. Expression vectors for these cells can include expression control sequences, such as an origin of replication, a promoter, an enhancer (Queen *et al.* (1986) Immunol. Rev. 89: 49), and necessary processing information sites, such as ribosome binding sites, RNA splice sites, polyadenylation sites, and transcriptional terminator sequences. Preferred expression control sequences are promoters derived from immunoglobulin genes, cytomegalovirus, SV40, Adenovirus, Bovine Papilloma Virus, and the like.

Eukaryotic DNA transcription can be increased by inserting an enhancer sequence into the vector. Enhancers are cis-acting sequences of between 10 to 300 bp that increase transcription by a promoter. Enhancers can effectively increase transcription when either 5' or 3' to the transcription unit. They are also effective if located within an intron or within the coding sequence itself. Typically, viral enhancers are used, including SV40 enhancers, cytomegalovirus enhancers, polyoma enhancers, and adenovirus enhancers. Enhancer sequences from mammalian systems are also commonly used, such as the mouse immunoglobulin heavy chain enhancer.

Mammalian expression vector systems will also typically include a selectable marker gene. Examples of suitable markers include, the dihydrofolate reductase gene (DHFR), the thymidine kinase gene (TK), or prokaryotic genes conferring drug resistance. The first two marker genes prefer the use of mutant cell lines that lack the ability to grow without the addition of thymidine to the growth medium. Transformed cells can then be identified by their ability to grow on non-supplemented media. Examples of prokaryotic drug resistance genes useful as markers include genes conferring resistance to G418, mycophenolic acid and hygromycin.

The vectors containing the DNA segments of interest can be transferred into the host

cell by well-known methods, depending on the type of cellular host. For example, calcium chloride transfection is commonly utilized for prokaryotic cells, whereas calcium phosphate treatment, lipofection, or electroporation may be used for other cellular hosts. Other methods used to transform mammalian cells include the use of Polybrene, protoplast fusion, liposomes, electroporation, and micro-injection (see, generally, Sambrook et al., supra).

Once expressed, the antibodies, individual mutated immunoglobulin chains, mutated antibody fragments, and other immunoglobulin polypeptides of the invention can be purified according to standard procedures of the art, including ammonium sulfate precipitation, fraction column chromatography, gel electrophoresis and the like (see, generally, Scopes, R., Protein Purification, Springer-Verlag, N.Y. (1982)). once purified, partially or to homogeneity as desired, the polypeptides may then be used therapeutically or in developing and performing assay procedures, immunofluorescent stainings, and the like (see, generally, Immunological Methods, Vols. I and II, Eds. Lefkovits and Pernis, Academic Press, New York, N.Y. (1979 and 1981)).

The antibodies generated by the method of the present invention can be used for diagnosis and therapy. By way of illustration and not limitation, they can be used to treat cancer, autoimmune diseases, or viral infections. For treatment of cancer, the antibodies will typically bind to an antigen expressed preferentially on cancer cells, such as erbB-2, CEA, CD33, and many other antigens and binding members well known to those skilled in the art.

4.16.5.9 Yeast Two-Hybrid Screening Assays

Shuffling can also be used to recombinatorially diversify a pool of selected library members obtained by screening a two-hybrid screening system to identify library members which bind a predetermined polypeptide sequence. The selected library members are pooled and shuffled by *in vitro* and/or *in vivo* recombination. The shuffled pool can then be screened in a yeast two hybrid system to select library

members which bind said predetermined polypeptide sequence (e.g., an SH2 domain) or which bind an alternate predetermined polypeptide sequence (e.g., an SH2 domain from another protein species).

An approach to identifying polypeptide sequences which bind to a predetermined polypeptide sequence has been to use a so-called "two-hybrid" system wherein the predetermined polypeptide sequence is present in a fusion protein (Chien *et al.* (1991) Proc. Natl. Acad. Sci. (USA) 88: 9578). This approach identifies protein-protein interactions *in vivo* through reconstitution of a transcriptional activator (Fields S and Song O (1989) Nature 340: 245), the yeast Gal4 transcription protein. Typically, the method is based on the properties of the yeast Gal4 protein, which consists of separable domains responsible for DNA-binding and transcriptional activation. Polynucleotides encoding two hybrid proteins, one consisting of the yeast Gal4 DNA-binding domain fused to a polypeptide sequence of a known protein and the other consisting of the Gal4 activation domain fused to a polypeptide sequence of a second protein, are constructed and introduced into a yeast host cell. Intermolecular binding between the two fusion proteins reconstitutes the Gal4 DNA-binding domain with the Gal4 activation domain, which leads to the transcriptional activation of a reporter gene (e.g., *lacZ*, *HIS3*) which is operably linked to a Gal4 binding site. Typically, the two-hybrid method is used to identify novel polypeptide sequences which interact with a known protein (Silver SC and Hunt SW (1993) Mol. Biol. Rep. 17: 155; Durfee *et al.* (1993) Genes Devel. 7: 555; Yang *et al.* (1992) Science 257: 680; Luban *et al.* (1993) Cell 73: 1067; Hardy *et al.* (1992) Genes Devel. 6: 801; Bartel *et al.* (1993) Biotechniques 14: 920; and Vojtek *et al.* (1993) Cell 74: 205). However, variations of the two-hybrid method have been used to identify mutations of a known protein that affect its binding to a second known protein (Li B and Fields S (1993) FASEB J. 7: 957; Lalo *et al.* (1993) Proc. Natl. Acad. Sci. (USA) 90: 5524; Jackson *et al.* (1993) Mol. Cell. Biol. 13: 2899; and Madura *et al.* (1993) J. Biol. Chem. 268: 12046). Two-hybrid systems have also been used to identify interacting structural domains of two known proteins (Bardwell *et al.* (1993) med. Microbial. 8: 1177; Chakrabarty *et al.* (1992) J. Biol. Chem. 267: 17498; Staudinger *et al.* (1993) J. Biol.

Chem. 268: 4608; and Milne GT. and Weaver DT (1993) Genes Devel. 7; 1755) or domains responsible for oligomerization of a single protein (Iwabuchi *et al.* (1993) Oncogene 8; 1693; Bogerd *et al.* (1993) J. Virol. 67: 5030). Variations of two-hybrid systems have been used to study the *in vivo* activity of a proteolytic enzyme (Dasmahapatra *et al.* (1992) Proc. Natl. Acad. Sci. (USA) 89: 4159). Alternatively, an *E. coli*/BCCP interactive screening system (Germino *et al.* (1993) Proc. Natl. Acad. Sci. (U.S.A.) 90: 933; Guarente L (1993) Proc. Natl. Acad. Sci. (U.S.A.) 90: 1639) can be used to identify interacting protein sequences (*i.e.*, protein sequences which heterodimerize or form higher order heteromultimers). Sequences selected by a two-hybrid system can be pooled and shuffled and introduced into a two-hybrid system for one or more subsequent rounds of screening to identify polypeptide sequences which bind to the hybrid containing the predetermined binding sequence. The sequences thus identified can be compared to identify consensus sequence(s) and consensus sequence kernels.

In general, standard techniques of recombination DNA technology are described in various publications, *e.g.* Sambrook *et al.*, 1989, Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory; Ausubel *et al.*, 1987, Current Protocols in Molecular Biology, vols. 1 and 2 and supplements, and Berger and Kimmel, Methods in Enzymology, Volume 152, Guide to Molecular Cloning Techniques (1987), Academic Press, Inc., San Diego, CA, each of which is incorporated herein in their entirety by reference. Polynucleotide modifying enzymes were used according to the manufacturers recommendations. Oligonucleotides were synthesized on an Applied Biosystems Inc. Model 394 DNA synthesizer using ABI chemicals. If desired, PCR amplifiers for amplifying a predetermined DNA sequence may be selected at the discretion of the practitioner.

4.16.5.9.1 Formation of Dimers

One microgram samples of template DNA are obtained and treated with U.V. light to

cause the formation of dimers, including TT dimers, particularly purine dimers. U.V. exposure is limited so that only a few photoproducts are generated per gene on the template DNA sample. Multiple samples are treated with U.V. light for varying periods of time to obtain template DNA samples with varying numbers of dimers from U.V. exposure.

4.16.5.9.2 Random Priming Kit

A random priming kit which utilizes a non-proofreading polymerase (for example, Prime-It II Random Primer Labeling kit by Stratagene Cloning Systems) is utilized to generate different size polynucleotides by priming at random sites on templates which are prepared by U.V. light (as described above) and extending along the templates. The priming protocols such as described in the Prime-It II Random Primer Labeling kit may be utilized to extend the primers. The dimers formed by U.V. exposure serve as a roadblock for the extension by the non-proofreading polymerase. Thus, a pool of random size polynucleotides is present after extension with the random primers is finished.

8.16.5.9.3 Generation of a Selected Mutant Polynucleotide Sequence

The present invention is further directed to a method for generating a selected mutant polynucleotide sequence (or a population of selected polynucleotide sequences) typically in the form of amplified and/or cloned polynucleotides, whereby the selected polynucleotide sequence(s) possess at least one desired phenotypic characteristic (e.g., encodes a polypeptide, promotes transcription of linked polynucleotides, binds a protein, and the like) which can be selected for. One method for identifying hybrid polypeptides that possess a desired structure or functional property, such as binding to a predetermined biological macromolecule (e.g., a receptor), involves the screening of a large library of polypeptides for individual library members which possess the desired structure or functional property conferred by the amino acid sequence of the polypeptide.

4.16.5.9.4 Generating Libraries Suitable for Affinity Interaction Screening or Phenotypic Screening

In one embodiment, the present invention provides a method for generating libraries of displayed polypeptides or displayed antibodies suitable for affinity interaction screening or phenotypic screening. The method comprises (1) obtaining a first plurality of selected library members comprising a displayed polypeptide or displayed antibody and an associated polynucleotide encoding said displayed polypeptide or displayed antibody, and obtaining said associated polynucleotides or copies thereof wherein said associated polynucleotides comprise a region of substantially identical sequences, optimally introducing mutations into said polynucleotides or copies, (2) pooling the polynucleotides or copies, (3) producing smaller or shorter polynucleotides by interrupting a random or particularized priming and synthesis process or an amplification process, and (4) performing amplification, preferably PCR amplification, and optionally mutagenesis to homologously recombine the newly synthesized polynucleotides.

4.16.5.9.5 Producing Hybrid Polynucleotides Which Express a Useful Hybrid Polypeptide

It is a particularly preferred object of the invention to provide a process for producing hybrid polynucleotides which express a useful hybrid polypeptide by a series of steps comprising:

- (a) producing polynucleotides by interrupting a polynucleotide amplification or synthesis process with a means for blocking or interrupting the amplification or synthesis process and thus providing a plurality of smaller or shorter polynucleotides due to the replication of the polynucleotide being in various stages of completion;
- (b) adding to the resultant population of single- or double-stranded polynucleotides one or more single- or double-stranded oligonucleotides, wherein said

added oligonucleotides comprise an area of identity in an area of heterology to one or more of the single- or double-stranded polynucleotides of the population;

- (c) denaturing the resulting single- or double-stranded oligonucleotides to produce a mixture of single-stranded polynucleotides, optionally separating the shorter or smaller polynucleotides into pools of polynucleotides having various lengths and further optionally subjecting said polynucleotides to a PCR procedure to amplify one or more oligonucleotides comprised by at least one of said polynucleotide pools;
- (d) incubating a plurality of said polynucleotides or at least one pool of said polynucleotides with a polymerase under conditions which result in annealing of said single-stranded polynucleotides at regions of identity between the single-stranded polynucleotides and thus forming of a mutagenized double-stranded polynucleotide chain;
- (e) optionally repeating steps (c) and (d);
- (f) expressing at least one hybrid polypeptide from said polynucleotide chain, or chains; and
- (g) screening said at least one hybrid polypeptide for a useful activity.

In a preferred aspect of the invention, the means for blocking or interrupting the amplification or synthesis process is by utilization of uv light, DNA adducts, DNA binding proteins.

In one embodiment of the invention, the DNA adducts, or polynucleotides comprising the DNA adducts, are removed from the polynucleotides or polynucleotide pool, such as by a process including heating the solution comprising the DNA fragments prior to further processing.

Having thus disclosed exemplary embodiments of the present invention, it should be noted by those skilled in the art that the disclosures are exemplary only and that various other alternatives, adaptations and modifications may be made within the

scope of the present invention. Accordingly, the present invention is not limited to the specific embodiments as illustrated herein.

Without further elaboration, it is believed that one skilled in the art can, using the preceding description, utilize the present invention to its fullest extent. The following examples are to be considered illustrative and thus are not limiting of the remainder of the disclosure in any way whatsoever.

5. Engineering goals

5.1. General overview: successive cycles of recombination and screening/selection

The invention provides methods for artificially evolving cells to acquire a new or improved property by recursive sequence recombination. Briefly, recursive sequence recombination entails successive cycles of recombination to generate molecular diversity and screening/selection to take advantage of that molecular diversity. That is, a family of nucleic acid molecules is created showing substantial sequence and/or structural identity but differing as to the presence of mutations. These sequences are then recombined in any of the described formats so as to optimize the diversity of mutant combinations represented in the resulting recombinant library. Typically, any resulting recombinant nucleic acids or genomes are recursively recombined for one or more cycles of recombination to increase the diversity of resulting products. After this recursive recombination procedure, the final resulting products are screened and/or selected for a desired trait or property.

Alternatively, each recombination cycle can be followed by at least one cycle of screening or selection for molecules having a desired characteristic. In this embodiment, the molecule(s) selected in one round form the starting materials for generating diversity in the next round. The cells to be evolved can be bacteria, archaebacteria, or eukaryotic cells and can constitute a homogeneous cell line or mixed culture. Suitable cells for evolution include the bacterial and eukaryotic, cell lines commonly used in genetic engineering, protein expression, or the industrial production or conversion of proteins, enzymes, primary metabolites, secondary metabolites, fine, specialty or commodity chemicals. Suitable mammalian cells include those from, e.g., mouse, rat, hamster, primate, and human, both cell lines and primary cultures. Such cells include stem cells, including embryonic stem cells and hemopoietic stem cells, zygotes, fibroblasts, lymphocytes, Chinese hamster ovary (CHO), mouse fibroblasts (NIHM), kidney, liver, muscle, and skin cells. Other eukaryotic cells of interest include plant cells, such as maize, rice, wheat, cotton, soybean, sugarcane, tobacco, and arabidopsis; fish, algae, fungi (penicillium, aspergillus, podospora, neurospora, saccharomyces), insect (e.g.,

baculo lepidoptera), yeast (picchia and saccharomyces, Schizosaccharomyces pombe). Also of interest are many bacterial cell types, both gram- negative and gram-positive, such as *Bacillus subtilis*, *B. licehniformis*, *B. cereus*, *Escherichia coli*, *Streptomyces*, *Pseudomonas*, *Salmonella*, *Actinomycetes*, *Lactobacillus*, *Acelonitcbacter*, *Deinococcus*, and *Erwinia*. The complete genome sequences of *E. coli* and *Bacillus subtilis* are described by Blattner et al., *Science* 277, 1454- 1462 (1997); Kunst et al., *Nature* 390, 249-256 (1997).

5.2 Identification and development of new and/or improved drugs

The genomics revolution, by determining the DNA sequences of great numbers of genes from many different organisms, has considerably broadened the possibilities for drug discovery by identifying, large numbers of molecules that are potential targets of drug action. One area of drug development focusing upon generating new antimicrobial drugs. New antimicrobial drugs are needed to treat infections by drug resistant organisms, and new methods are urgently needed to facilitate making such discoveries. Technical advances in molecular biology, automated methods for high throughput screening and chemical syntheses have led to an increase in the number of target based screens utilized for antimicrobial drug discovery and in the number of compounds being analyzed.

The invention relates to procedures that can be applied to identifying compounds that bind to and modulate the function of target components of a cell whose function is known or unknown, and cell components that are not amenable to other screening methods. The invention relates to generating and/or identifying a compound that binds to and modulates (inhibits or enhances) the function of a component of a cell, thereby producing a phenotypic effect in the cell. Within these procedures are methods for identifying a biomolecule that 1) binds to, *in vitro*, a component of a cell that has been isolated from other constituents of the cell and that 2) causes, *in vivo*, as seen in an assay upon intracellular expression of the biomolecule, a phenotypic effect in the cell which is the usual producer and host of the target cell component. In an assay demonstrating

characteristic 2) above, intracellular production of the biomolecule can be in cells grown in culture or in cells introduced into an animal. Further methods within these procedures are those methods comprising an assay for a phenotypic effect in the cell upon intracellular production of the biomolecule, either in cells in culture or in cells that have been introduced into one or more animals, and an assay to identify one or more compounds that behave as competitors of the biomolecule in an assay of binding to the target cell component.

5.2.1. Procedure for identifying and/or designing compounds with antimicrobial activity against a pathogen

The invention further relates to methods particularly well suited to a procedure for identifying and/or designing compounds with antimicrobial activity against a pathogen whose target cell component is the subject of studies to identify such compounds. A common mechanism of action of an antimicrobial agent is binding to a component of the cells of the pathogen treated with the antimicrobial.

The procedure includes methods for identifying biomolecules that bind to a chosen target *in vitro*, methods for identifying biomolecules that also bind to the chosen target and modulate its function during infection of a host mammal *in vivo*, and methods for identifying compounds that compete with the biomolecules for sites on the target in competitive binding assays. Compounds identified by this procedure are candidates for drugs with antimicrobial activity against the pathogen.

5.3 Producing proteins with improved affinities

Polynucleotide sequences selected in a first selection round (typically by affinity selection for binding to a receptor (e.g., a ligand)) by any of these methods are pooled and the pool(s) is/are shuffled by *in vitro* and/or *in vivo* recombination to produce a shuffled pool comprising a population of recombinant polynucleotide sequences. The recombinant polynucleotide sequences are subjected to at least one subsequent selection round. The polynucleotide sequences selected in the subsequent selection round(s) can be used directly, sequenced, and/or subjected to one or more

additional rounds of shuffling and subsequent selection. Selected sequences can also be back-crossed with polynucleotide sequences encoding neutral sequences (i.e., having insubstantial functional effect on binding), such as for example by back-crossing with a wild-type or naturally-occurring sequence substantially identical to a selected sequence to produce native-like functional peptides, which may be less immunogenic. Generally, during back-crossing subsequent selection is applied to retain the property of binding to the predetermined receptor (ligand).

Prior to or concomitant with the shuffling of selected sequences, the sequences can be mutagenized. In one embodiment, selected library members are cloned in a prokaryotic vector (e.g., plasmid, phagemid, or bacteriophage) wherein a collection of individual colonies (or plaques) representing discrete library members are produced. Individual selected library members can then be manipulated (e.g., by site-directed mutagenesis, cassette mutagenesis, chemical mutagenesis, PCR mutagenesis, and the like) to generate a collection of library members representing a kernal of sequence diversity based on the sequence of the selected library member. The sequence of an individual selected library member or pool can be manipulated to incorporate random mutation, pseudorandom mutation, defined kernal mutation (i.e., comprising variant and invariant residue positions and/or comprising variant residue positions which can comprise a residue selected from a defined subset of amino acid residues), codon-based mutation, and the like, either segmentally or over the entire length of the individual selected library member sequence. The mutagenized selected library members are then shuffled by *in vitro* and/or *in vivo* recombinatorial shuffling as disclosed herein.

The invention also provides a product-by-process, wherein selected polynucleotide sequences having (or encoding a peptide having) a predetermined binding specificity are formed by the process of: (1) screening a displayed peptide or displayed single-chain antibody library against a predetermined receptor (e.g., ligand) or epitope (e.g., antigen macromolecule) and identifying and/or enriching library members which bind to the predetermined receptor or epitope to produce a pool of selected library members, (2) shuffling by recombination the selected library members (or amplified or cloned copies thereof) which binds the predetermined epitope and has been thereby isolated and/or enriched from the library to generate a shuffled library, and (3) screening

the shuffled library against the predetermined receptor (e.g., ligand) or epitope (e.g., antigen macromolecule) and identifying and/or enriching shuffled library members which bind to the predetermined receptor or epitope to produce a pool of selected shuffled library members.

In one embodiment, the present invention provides a method for generating libraries of displayed polypeptides or displayed antibodies suitable for affinity interaction screening or phenotypic screening. The method comprises (1) obtaining a first plurality of selected library members comprising a displayed polypeptide or displayed antibody and an associated polynucleotide encoding said displayed polypeptide or displayed antibody, and obtaining said associated polynucleotides or copies thereof wherein said associated polynucleotides comprise a region of substantially identical sequences, optimally introducing mutations into said polynucleotides or copies, (2) pooling the polynucleotides or copies, (3) producing smaller or shorter polynucleotides by interrupting a random or particularized priming and synthesis process or an amplification process, and (4) performing amplification, preferably PCR amplification, and optionally mutagenesis to homologously recombine the newly synthesized polynucleotides.

It is a particularly preferred object of the invention to provide a process for producing hybrid polynucleotides which express a useful hybrid polypeptide by a series of steps comprising:

- (a) producing polynucleotides by interrupting a polynucleotide amplification or synthesis process with a means for blocking or interrupting the amplification or synthesis process and thus providing a plurality of smaller or shorter polynucleotides due to the replication of the polynucleotide being in various stages of completion;
- (b) adding to the resultant population of single- or double-stranded polynucleotides one or more single- or double-stranded oligonucleotides, wherein said added oligonucleotides comprise an area of identity in an area of heterology to one or more of the single- or double-stranded polynucleotides of the population;
- (c) denaturing the resulting single- or double-stranded oligonucleotides to produce a mixture of single-stranded polynucleotides, optionally separating the shorter or smaller polynucleotides into pools of polynucleotides having various lengths and further

optionally subjecting said polynucleotides to a PCR procedure to amplify one or more oligonucleotides comprised by at least one of said polynucleotide pools;

(d) incubating a plurality of said polynucleotides or at least one pool of said polynucleotides with a polymerase under conditions which result in annealing of said single-stranded polynucleotides at regions of identity between the single-stranded polynucleotides and thus forming of a mutagenized double-stranded polynucleotide chain;

(e) optionally repeating steps (c) and (d);

(f) expressing at least one hybrid polypeptide from said polynucleotide chain, or chains; and

(g) screening said at least one hybrid polypeptide for a useful activity.

In a preferred aspect of the invention, the means for blocking or interrupting the amplification or synthesis process is by utilization of UV light, DNA adducts, DNA binding proteins.

In one embodiment of the invention, the DNA adducts, or polynucleotides comprising the DNA adducts, are removed from the polynucleotides or polynucleotide pool, such as by a process including heating the solution comprising the DNA fragments prior to further processing.

Having thus disclosed exemplary embodiments of the present invention, it should be noted by those skilled in the art that the disclosures are exemplary only and that various other alternatives, adaptations and modifications may be made within the scope of the present invention. Accordingly, the present invention is not limited to the specific embodiments as illustrated herein.

5.3.1. Antibody production

High affinity single-chain antibodies of the desired specificity can be engineered and expressed in a variety of systems. For example, scfv have been produced in plants (Firek et al, 1993) and can be readily made in prokaryotic systems (Owens and Young, 1994; Johnson and Bird, 1991). Furthermore, the single-chain antibodies can be used as a basis for constructing whole antibodies or various fragments thereof (Kettleborough et al, 1994). The variable region encoding sequence may be isolated (e.g., by PCR

amplification or subcloning) and spliced to a sequence encoding a desired human constant region to encode a human sequence antibody more suitable for human therapeutic uses where immunogenicity is preferably minimized. The polynucleotide(s) having the resultant fully human encoding sequence(s) can be expressed in a host cell (e.g., from an expression vector in a mammalian cell) and purified for pharmaceutical formulation. The antibodies generated by the method of the present invention can be used for diagnosis and therapy. By way of illustration and not limitation, they can be used to treat cancer, autoimmune diseases, or viral infections. For treatment of cancer, the antibodies will typically bind to an antigen expressed preferentially on cancer cells, such as erbB-2, CEA, CD33, and many other antigens and binding members well known to those skilled in the art.

5.3.1.1. Modified variable regions

Beginning in 1988, single-chain analogues of Fv fragments and their fusion proteins have been reliably generated by antibody engineering methods. The first step generally involves obtaining the genes encoding VH and VL domains with desired binding properties; these V genes may be isolated from a specific hybridoma cell line, selected from a combinatorial V-gene library, or made by V gene synthesis. The single-chain Fv is formed by connecting the component V genes with an oligonucleotide that encodes an appropriately designed linker peptide, such as (Gly-Gly-Gly-Gly-Ser)3 or equivalent linker peptide(s). The linker bridges the C-terminus of the first V region and N-terminus of the second, ordered as either VH-linker-VL or VL-linker-VH'. In principle, the scfv binding site can faithfully replicate both the affinity and specificity of its parent antibody combining site.

Thus, scfv fragments are comprised of VH and VL domains linked into a single polypeptide chain by a flexible linker peptide. After the scfv genes are assembled, they are cloned into a phagemid and expressed at the tip of the M13 phage (or similar filamentous bacteriophage) as fusion proteins with the bacteriophage PIII (gene 3) coat protein. Enriching for phage expressing an antibody of interest is accomplished by

panning the recombinant phage displaying a population scfv for binding to a predetermined epitope (e.g., target antigen, receptor).

The linked polynucleotide of a library member provides the basis for replication of the library member after a screening or selection procedure, and also provides the basis for the determination, by nucleotide sequencing, of the identity of the displayed peptide sequence or VH and VL amino acid sequence. The displayed peptide (s) or single-chain antibody (e.g., scfv) and/or its VH and VL domains or their CDRs can be cloned and expressed in a suitable expression system. Often polynucleotides encoding the isolated VH and VL domains will be ligated to polynucleotides encoding constant regions (CH and CL) to form polynucleotides encoding complete antibodies (e.g., chimeric or fully-human), antibody fragments, and the like. Often polynucleotides encoding the isolated CDRs will be grafted into polynucleotides encoding a suitable variable region framework (and optionally constant regions) to form polynucleotides encoding complete antibodies (e.g., humanized or fully-human), antibody fragments, and the like. Antibodies can be used to isolate preparative quantities of the antigen by immunoaffinity chromatography. Various other uses of such antibodies are to diagnose and/or stage disease (e.g., neoplasia) and for therapeutic application to treat disease, such as for example: neoplasia, autoimmune disease, AIDS, cardiovascular disease, infections, and the like.

If it were possible to generate scfv libraries having broader antibody diversity and overcoming many of the limitations of conventional CDR mutagenesis and randomization methods which can cover only a very tiny fraction of the potential sequence combinations, the number and quality of scfv antibodies suitable for therapeutic and diagnostic use could be vastly improved. To address this, the *in vitro* and *in vivo* shuffling methods of the invention are used to recombine CDRs which have been obtained (typically via PCR amplification or cloning) from nucleic acids obtained from selected displayed antibodies. Such displayed antibodies can be displayed on cells, on bacteriophage particles, on polysomes, or any suitable antibody display system wherein the antibody is associated with its encoding nucleic acid(s). In a variation, the CDRs are initially obtained from mRNA (or cDNA) from antibody-producing cells (e.g., plasma cells/splenocytes from an immunized wild-type mouse, a human, or a transgenic mouse

capable of making a human antibody as in **WO 92/03918**, **WO 93/12227**, and **WO 94/25585**), including hybridomas derived therefrom.

Polynucleotide sequences selected in a first selection round (typically by affinity selection for displayed antibody binding to an antigen (e.g., a ligand) by any of these methods are pooled and the pool(s) is/are shuffled by *in vitro* and/or *in vivo* recombination, especially shuffling of CDRs (typically shuffling heavy chain CDRs with other heavy chain CDRs and light chain CDRs with other light chain CDRs) to produce a shuffled pool comprising a population of recombined selected polynucleotide sequences. The recombined selected polynucleotide sequences are expressed in a selection format as a displayed antibody and subjected to at least one subsequent selection round. The polynucleotide sequences selected in the subsequent selection round(s) can be used directly, sequenced, and/or subjected to one or more additional rounds of shuffling and subsequent selection until an antibody of the desired binding affinity is obtained. Selected sequences can also be back-crossed with polynucleotide sequences encoding neutral antibody framework sequences (i.e., having insubstantial functional effect on antigen binding), such as for example by back-crossing with a human variable region framework to produce human-like sequence antibodies. Generally, during back-crossing subsequent selection is applied to retain the property of binding to the predetermined antigen.

Alternatively, or in combination with the noted variations, the valency of the target epitope may be varied to control the average binding affinity of selected scfv library members. The target epitope can be bound to a surface or substrate at varying densities, such as by including a competitor epitope, by dilution, or by other method known to those in the art. A high density (valency) of predetermined epitope can be used to enrich for scfv library members which have relatively low affinity, whereas a low density (valency) can preferentially enrich for higher affinity scfv library members.

For generating diverse variable segments, a collection of synthetic oligonucleotides encoding random, pseudorandom, or a defined sequence kernel set of peptide sequences can be inserted by ligation into a predetermined site (e.g., a CDR). Similarly, the sequence diversity of one or more CDRs of the single-chain antibody

cassette(s) can be expanded by mutating the CDR(s) with site-directed mutagenesis, CDR-replacement, and the like. The resultant DNA molecules can be propagated in a host for cloning and amplification prior to shuffling, or can be used directly (i.e., may avoid loss of diversity which may occur upon propagation in a host cell) and the selected library members subsequently shuffled.

A variety of techniques can be used in the present invention to diversify a peptide library or single-chain antibody library, or to diversify, prior to or concomitant with shuffling, around variable segment peptides found in early rounds of panning to have sufficient binding activity to the predetermined macromolecule or epitope. In one approach, the positive selected peptide/polynucleotide complexes (those identified in an early round of affinity enrichment) are sequenced to determine the identity of the active peptides. Oligonucleotides are then synthesized based on these active peptide sequences, employing a low level of all bases incorporated at each step to produce slight variations of the primary oligonucleotide sequences. This mixture of (slightly) degenerate oligonucleotides is then cloned into the variable segment sequences at the appropriate locations. This method produces systematic, controlled variations of the starting peptide sequences, which can then be shuffled. It requires, however, that individual positive nascent peptide/polynucleotide complexes be sequenced before mutagenesis, and thus is useful for expanding the diversity of small numbers of recovered complexes and selecting variants having higher binding affinity and/or higher binding specificity. In a variation, mutagenic PCR amplification of positive selected peptide/polynucleotide complexes (especially of the variable region sequences, the amplification products of which are shuffled *in vitro* and/or *in vivo* and one or more additional rounds of screening is done prior to sequencing. The same general approach can be employed with single-chain antibodies in order to expand the diversity and enhance the binding affinity/specification, typically by diversifying CDRs or adjacent framework regions prior to or concomitant with shuffling. If desired, shuffling reactions can be spiked with mutagenic oligonucleotides capable of *in vitro* recombination with the selected library members can be included. Thus, mixtures of synthetic oligonucleotides and PCR produced polynucleotides (synthesized by error-prone or high-fidelity methods) can be added to the

in vitro shuffling mix and be incorporated into resulting shuffled library members (shufflants).

5.3.1.2. Modified CDR regions

The present invention of shuffling enables the generation of a vast library of CDR-variant single-chain antibodies. One way to generate such antibodies is to insert synthetic CDRs into the single-chain antibody and/or CDR randomization prior to or concomitant with shuffling. The sequences of the synthetic CDR cassettes are selected by referring to known sequence data of human CDR and are selected in the discretion of the practitioner according to the following guidelines: synthetic CDRs will have at least 40 percent positional sequence identity to known CDR sequences, and preferably will have at least 50 to 70 percent positional sequence identity to known CDR sequences. For example, a collection of synthetic CDR sequences can be generated by synthesizing a collection of oligonucleotide sequences on the basis of naturally-occurring human CDR sequences listed in Kabat (Kabat et al, 1991); the pool (s) of synthetic CDR sequences are calculated to encode CDR peptide sequences having at least 40 percent sequence identity to at least one known naturally-occurring human CDR sequence. Alternatively, a collection of naturally-occurring CDR sequences may be compared to generate consensus sequences so that amino acids used at a residue position frequently (i.e., in at least 5 percent of known CDR sequences) are incorporated into the synthetic CDRs at the corresponding position(s). Typically, several (e.g., 3 to about 50) known CDR sequences are compared and observed natural sequence variations between the known CDRs are tabulated, and a collection of oligonucleotides encoding CDR peptide sequences encompassing all or most permutations of the observed natural sequence variations is synthesized. For example but not for limitation, if a collection of human VH CDR sequences have carboxy-terminal amino acids which are either Tyr, Val, Phe, or Asp, then the pool(s) of synthetic CDR oligonucleotide sequences are designed to allow the carboxy-terminal CDR residue to be any of these amino acids. In some embodiments, residues other than those which naturally-occur at a residue position in the collection of CDR sequences are incorporated: conservative amino acid substitutions are frequently incorporated and up to 5 residue positions may be varied to incorporate non-conservative

amino acid substitutions as compared to known naturally-occurring CDR sequences. Such CDR sequences can be used in primary library members (prior to first round screening) and/or can be used to spike *in vitro* shuffling reactions of selected library member sequences. Construction of such pools of defined and/or degenerate sequences will be readily accomplished by those of ordinary skill in the art.

The collection of synthetic CDR sequences comprises at least one member that is not known to be a naturally-occurring CDR sequence. It is within the discretion of the practitioner to include or not include a portion of random or pseudorandom sequence corresponding to N region addition in the heavy chain CDR; the N region sequence ranges from 1 nucleotide to about 4 nucleotides occurring at V-D and D-J junctions. A collection of synthetic heavy chain CDR sequences comprises at least about 100 unique CDR sequences, typically at least about 1,000 unique CDR sequences, preferably at least about 10,000 unique CDR sequences, frequently more than 50,000 unique CDR sequences; however, usually not more than about 1×10^6 unique CDR sequences are included in the collection, although occasionally 1×10^7 to 1×10^8 unique CDR sequences are present, especially if conservative amino acid substitutions are permitted at positions where the conservative amino acid substituent is not present or is rare (i.e., less than 0.1 percent) in that position in naturally-occurring human CDRS. In general, the number of unique CDR sequences included in a library should not exceed the expected number of primary transformants in the library by more than a factor of 10. Such single-chain antibodies generally bind of about at least 1×10^{-7} M⁻¹, preferably with an affinity of about at least 5×10^7 M⁻¹, more preferably with an affinity of at least 1×10^8 M⁻¹ to 1×10^9 M⁻¹ or more, sometimes up to 1×10^{10} M⁻¹ or more. Frequently, the predetermined antigen is a human protein, such as for example a human cell surface antigen (e.g., CD4, CD8, IL-2 receptor, EGF receptor, PDGF receptor), other human biological macromolecule (e.g., thrombomodulin, protein C, carbohydrate antigen, sialyl Lewis antigen, Lselectin), or nonhuman disease associated macromolecule (e.g., bacterial LPS, virion capsid protein or envelope glycoprotein) and the like.

5.4 Increased expression in a recombinant host

In one embodiment of this invention, it provides for increasing expression of a gene or trait of interest in a recombinant host of interest. The hosts can include but are not limited to bacteria, fungi, protozoans, viruses, animals, insects, and plants.

5.5 Metabolite shifting

In one embodiment of this invention, it provides for metabolite shifting.

5.6 Creating a modified plant with desired traits

One aspect of the present invention relates to the use of trait DNA molecules which are heterologous to the plant -- e.g., DNA molecules that confer disease resistance to plants transformed with the DNA construct. The present invention is useful in plants for imparting resistance to a wide variety of pathogens including viruses, bacteria, fungi, viroids, phytoplasmas, nematodes, and insects. The present invention may also be used in mammals to impart genetic traits. Resistance, *inter alia*, to the following viruses can be achieved by the method of the present invention: tomato spotted wilt virus, impatiens necrotic spot virus, groundnut ringspot virus, potato virus Y, potato virus X, tobacco mosaic virus, turnip mosaic virus, tobacco etch virus, papaya ringspot virus, tomato mottle virus, tomato yellow leaf curl virus, or-combinations thereof. Resistance, *inter alia*, to the following bacteria can also be imparted to plants in accordance with present invention: *Pseudomonas solancearum*, *Pseudomonas syringae* pv. *tabaci*, *Xanthomonas campestris* pv. *pelargonii*, and *Agrobacterium tumefaciens*. Plants can be made resistant, *inter alia*, to the following fungi by use of the method of the present invention: *Fusarium oxysporum* and *Phytophthora infestans*. Suitable DNA molecules include a DNA molecule encoding a coat protein, a replicase, a DNA molecule not encoding protein, a DNA molecule encoding a viral gene product, or combinations thereof.

The present invention is also used to confer traits other than disease resistance on plants. For example, DNA molecules which impart a plant genetic trait can be used as the DNA trait molecule of the present invention. In this aspect of the present invention, suitable trait DNA molecules encode for desired color, enzyme production, or combinations thereof. In another embodiment of this invention, it provides for

engineering plants with desired traits, including output (e.g., producing increased amounts of a desired vitamin, mineral, or genetically engineered and introduced molecule such as an antibody) and input (e.g., drought and/or salinity resistance) traits.

5.7 PLANT GENE EXPRESSION

This invention is related to the genetic engineering of plants and to a means and method (use of DNA construct) for conferring a plurality of traits, including resistance to viruses, to a plant using a vector encoding a plurality of genes, such as coat protein genes, protease genes, or replicase genes. The field of the invention is plant genetics, including genetic mapping and restriction fragment length polymorphism technology.

The present invention also relates to:

- (i) the production of mature proteins in plant cells, including the production of proteins in mature secreted form.
- (ii) the development of techniques for the commercial production of transgenic plants.

5.7.1 General Considerations

The present invention provides a chimeric recombinant DNA molecule comprising: a plurality of DNA sequences, each of which comprises a plant-functional promoter linked to a coding region, which encodes a virus-associated coat protein, wherein said DNA sequences are preferably linked in-tandem so that they are expressed in virus- susceptible plant cells transformed with said recombinant DNA molecule to impart resistance to said viruses; as well as methods for transforming plants with the chimeric constructs and for selecting plants which express at least one of said DNA sequences imparting viral resistance.

Methods of making a genetically modified plant comprising regenerating a whole plant from a plant cell that has been transfected with DNA sequences comprising a first gene whose expression results in an altered plant phenotype linked to a transiently active

promoter, the gene and promoter being separated by a blocking sequence flanked on either side by specific excision sequences, a second gene that encodes a recombinase specific for the specific excision sequences linked to a repressible promoter, and a third gene that encodes the repressor specific for the repressible promoter. Also a method for making a genetically modified hybrid plant by hybridizing a first plant regenerated from a plant cell that has been transfected with DNA sequences comprising a first gene whose expression results in an altered plant phenotype linked to a transiently active promoter, the gene and promoter being separated by a blocking sequence flanked on either side by specific excision sequences to a second plant regenerated from a second plant cell that has been transfected with DNA sequences comprising a second gene that encodes a recombinase specific for the specific excision sequences linked to a promoter that is active during seed germination, and growing a hybrid plant from the hybrid seed. Plant cells, plant tissues, plant seed and whole plants containing the above DNA sequences are also claimed.

The present invention is also directed to a DNA construct formed from a fusion gene which includes a trait DNA molecule and a silencer DNA molecule. The trait DNA molecule has a length that is insufficient to impart a desired trait to plants transformed with the trait DNA molecule. The silencer DNA molecule is operatively coupled to the trait DNA molecule with the trait and silencer DNA molecules collectively having sufficient length to impart the trait to plants transformed with the DNA construct. Expression systems, host cells, plants, and plant seeds containing the DNA construct are disclosed. The present invention is also directed to imparting multiple traits to a plant.

The present invention is also directed to methods of introgressing one or more desired quantitative traits into a plant comprising screening one or more restriction fragment length polymorphisms (RFLP) for association with desired quantitative traits (QT), selecting one or more RFLP's showing association with the desired QT's, developing a mathematical model based on the magnitude of the association of RFLP(s) to predict the degree of expression of the desired QT's, and using the thus-selected RFLP(s) and the mathematical model in a plant breeding program to predict the degree of introgression and expression of the desired QT's in plant progeny.

A method for producing one of the following proteins in transgenic monocot plant cells is disclosed: (i) mature, glycosylated α_1 -antitrypsin (AAT) having the same N-terminal amino acid sequence as mature AAT produced in humans and a glycosylation pattern which increases serum halflife substantially over that of mature non-glycosylated AAT; (ii) mature, glycosylated antithrombin III (ATIII) having the same N-terminal amino acid sequence as mature ATIII produced in humans; (iii) mature human serum albumin (HSA) having the same N-terminal amino acid sequence as mature HSA produced in humans and having the folding pattern of native mature HSA as evidenced by its bilirubin- binding characteristics; and (iv) mature, active subtilisin BPN' (BPN') having the same N-terminal amino acid sequence as BPN' produced in *Bacillus*.

Monocot plants cells are transformed with a chimeric gene which includes a DNA coding sequence encoding a fusion protein having an (i) N-terminal moiety corresponding to a rice α -amylase signal sequence peptide and, (iii) immediately adjacent the C-terminal amino acid of said peptide, a protein moiety corresponding to the mature protein to be produced.

A process for commercially propagating plants by tissue culture in such a way as both to conserve desired plant morphology and to transform the plant with respect to one or more desired genes. The method includes the steps of (a) creating an Agrobacterium vector containing the gene sequence desired to be transferred to the propagated plant, preferably together with a marker gene; (b) taking one or more petiole explants from a mother plant and inoculating them with the Agrobacterium vector; (c) conducting callus formation in the petiole sections in culture, in the dark; and (d) culturing the resulting callus in growth medium containing a benzylamino growth regulator such as benzylaminopurine or, most preferably, benzylaminopurine-riboside. Additional optional growth regulators including auxins and cytokinins (indole butyric acid, benzylamine, benzyladenine, benzylaminopurine, alpha naphthylacetic acid and others known in the art) may also be present. Preferably, the petiole tissue is taken from *Pelargonium x domesticum* and the Agrobacterium vector contains an antisense gene for ACC synthase or ACC oxidase to prevent ACC synthase or ACC oxidase expression and, in turn, the ethylene formation for which these enzymes are precursors.

5.7.2. PRODUCTION OF VIRUS RESISTANT PLANTS

5.7.2.1 Production of virus resistant plants

Scientists have recently developed means to produce virus resistant plants using genetic engineering techniques. Several different types of host resistance to viruses are recognized. The host may be resistant to: (1) establishment of infection, (2) virus multiplication, or (3) viral movement. One potential application would be to engineer a plant that is resistant to potyviruses. Potyviruses are a distinct group of plant viruses which are pathogenic to various crops, and which demonstrate cross-infectivity between plant members of different families. One example is that expression of the coat protein genes from tobacco mosaic virus, alfalfa mosaic virus, cucumber mosaic virus, and potato virus X, among others, in transgenic plants has resulted in plants which are resistant to infection by the respective virus. Some evidence of heterologous protection has also been reported. For example, see references Namba et al., Phytopathology, 82, 940 (1992) Stark et al., Biotechnology, 1, 1257 (1989).

5.7.2.2 USING "PATHOGEN DRIVEN RESISTANCE" (PDR) FOR DEVELOPING VIRUS RESISTANT TRANSGENIC PLANTS

Control of plant virus diseases took a major step forward in the last decade when it was shown in 1986 that the tobacco mosaic virus ("TMV") coat protein gene that was expressed in transgenic tobacco conferred resistance to TMV (Powell-Abel, P., et al., "Delay of Disease Development in Transgenic Plants that Express the Tobacco Mosaic Virus Coat Protein Gene," Science, 232:738-43 (1986)). The concept of pathogen-derived resistance ("PDR"), which states that pathogen genes that are expressed in transgenic plants will confer resistance to infection by the homologous or related pathogens (Sanford, J.C., et al. "The Concept of Parasite-Derived Resistance - Deriving Resistance Genes from the Parasite's Own Genome," J. Theor. Biol., 113:395-405 (1985)) was introduced at about the same time. Since then, numerous reports have confirmed that PDR is a useful strategy for developing transgenic plants that are resistant to many

different viruses (Lomonosoff, G.P., "Pathogen-Derived Resistance to Plant Viruses," Ann. Rev. Photopathol., 33:323-43 (1995)).

Only eight years after the report by Beachy and colleagues (Powell-Abel, P., et al., "Delay of Disease Development in Transgenic Plants that Express the Tobacco Mosaic Virus Coat Protein Gene," Science, 232:738-43 (1986)), Grumet, R., "Development of Virus Resistant Plants via Genetic Engineering," Plant Breeding Reviews, 12:47-49 (1994) reviewed the PDR literature and listed the successful development of virus resistant transgenic plants to at least 11 different groups of plant viruses.

5.7.2.2.1 Utilizing The Coat Protein Genes

The vast majority of reports have utilized the coat protein genes of the viruses that are targeted for control (e.g., Grumet, R., "Development of Virus Resistant Plants via Genetic Engineering," Plant Breeding Reviews, 12:47-49 (1994)). Additional examples are included in the following references: Fuchs, M., et al., Bio/Technology, 13:1466-73 (1995); Tricoli, D.M., et al., Bio/Technology, 13:1458-65 (1995); Fitch, M. M. M., et al., Bio/Technology, 10:1466-72 (1992); Tennant, P.F., et al., Phytopathology, 84:1359-66 (1994).

5.7.2.2.2 Other effective viral genes

Interestingly, remarkable progress has been made in developing virus resistant transgenic plants despite a poor understanding of the mechanisms involved in the various forms of pathogen-derived resistance (Lomonosoff, G.P., "Pathogen-Derived Resistance to Plant Viruses," Ann.-Rev. Photopathol., 33:323-43 (1995)). Various reports have utilized proteins other than the coat protein genes to confer resistance (Golemboski, D.B., et al., Proc. Natl. Acad. Sci. USA, 87:6311-15 (1990); Beck, D. L., et al., Proc. Natl. Acad. Sci. USA, 91:10310-14 (1994); Maiti, I.B., et al., Proc. Natl. Acad. Sci. USA, 90:6110-14 (1993). Furthermore, the viral genes can be effective in the translatable and nontranslatable sense forms, and less frequently antisense forms (e.g., Baulcombe, D.C., "Mechanisms of Pathogen-Derived Resistance to Viruses in Transgenic Plants," Plant

Cell, 8:1833-44 (1996); Dougherty, W. G., et al., "Transgenes and Gene Suppression: Telling us Something New?," Current Opinion in Cell Biology, 7:399- 05 (1995); Lomonossoff, G.P., "Pathogen-Derived Resistance to Plant Viruses," Ann. Rev. Photopathol. 33:323-43 (1995)).

5.7.2.2.3. RNA-MEDIATED RESISTANCE

5.7. 2.2.3.1 Description (A Form of PDR)

RNA-mediated resistance is the form of PDR where there is clear evidence that viral proteins do not play a role in conferring resistance to the transgenic plant. The first clear cases for RNA-mediated resistance were reported in 1992 for tobacco etch ("TEV") potyvirus (Lindbo, et al., "Pathogen-Derived Resistance to a Potyvirus Immune and Resistance Phenotypes in Transgenic Tobacco Expressing Altered Forms of a Potyvirus Coat Protein Nucleotide Sequence," Mol. Plant Microbe Interact., 5:144-53 (1992)), for potato virus Y ("PVY") potyvirus by Van Der Vlugt, R.A.A., et al., "Evidence for Sense RNA-Mediated Protection to PVY in Tobacco Plants Transformed with the Viral Oat Protein Cistron," Plant Mol. Biol., 20:631-39 (1992), and for tomato spotted wilt ("TSWV") tospovirus by de Haan, P., et al., "Characterization of RNA-Mediated Resistance to Tomato Spotted Wilt Virus in Transgenic Tobacco Plants," Bio/Technology, 10:1133-37 (1992). others confirmed the occurrence of RNA-mediated resistance with potyviruses (Smith, H.A., et al., "Transgenic Plant Virus Resistance Mediated by Untranslatable Sense RNAs: Expression, Regulation, and Fate of Nonessential RNAs," Plant Cell, 6:1441-53 (1994)), potexviruses (Mueller, E., et al., "Homology-Dependent Resistance: Transgenic Virus Resistance in Plants Related to Homology-Dependent Gene Silencing," Plant Journal, 7:1001-13 (1995)), and TSWV and other topsoviruses (Pang, S.Z., et al., "Resistance of Transgenic Nicotiana Benthamiana Plants to Tomato Spotted Wilt and Impatiens Necrotic Spot Tospoviruses: Evidence of Involvement of the N Protein and N Gene RNA in Resistance," Phytopathology, 84:243-49 (1994); Pang, S.-Z., et al., "Different Mechanisms Protect Transgenic Tobacco Against Tomato Spotted Wilt Virus and Impatiens Necrotic Spot Tospoviruses," Bio/Technology 11:819-24 (1993)). More recent work has shown that RNA-mediated resistance also occurs with the

comovirus cowpea mosaic virus (Sijen, T., et al., "RNA-Mediated Virus Resistance: Role of Repeated Transgene and Delineation of Targeted Regions," *Plant Cell*, 8:2227-94 (1996)) and squash mosaic virus (Jan, F.-J., et al., "Genetic and Molecular Analysis of Squash Plants Transformed with Coat Protein Genes of Squash Mosaic Virus," *Phytopathology*, 86:S16-17 (1996)).

5.7.3 CREATING TRANSGENIC PLANTS WITH CONTROLLABLE GENES

This invention also relates to certain transgenic plants and involves a method of creating transgenic plants with controllable genes. More particularly, the invention relates to transgenic plants that have been modified such that expression of a desired introduced gene can be limited to a particular stage of plant development, a particular plant tissue, particular environmental conditions, or a particular time or location, or a combination of these situations.

5.7.3.1 Inducible gene promoter: "gene switch"

Various gene expression control elements that are operable in one or more species of organisms are known. Examples are mentioned in PCT Application WO 90/08826 and PCT application WO 94/03619. A Tetracycline-controlled plant-active repressor-operator system can be utilized as described in various references: Gatz and Quail (1988) and Gatz, et al. (1992), (Höpke-Seyler), 372:659-660 (1991); Gatz and Quail, 1988; and (Gatz, et al., 1992).

5.7.4 RECOMBINANT PRODUCTION OF PROTEINS

A major commercial focus of biotechnology is the recombinant production of proteins, including both industrial enzymes and proteins that have important therapeutic uses.

5.7.4.1. Alternative protein expression system to overcome problems of microbial and mammalian systems

It would therefore be desirable to produce selected therapeutic and industrial proteins in a protein expression system that largely overcomes problems associated with microbial and mammalian-cell systems. In particular, production of the proteins should

allow large volume production at low cost, and yield properly processed and glycosylated proteins. The production system should also have a relatively stable genotype from generation to generation. These aims are achieved, in the present invention, for the therapeutic proteins AAT, HSA, and antithrombin III (ATIII), and the industrial enzyme subtilisin BPN'.

5.7.4.2. Uses

Various proteins of interest could be produced such as but not limited to:

- 1) Human α_1 -antitrypsin (AAT; Carrell, P., et al., Nature (1992) 298:329; involved in cirrhosis and liver failure: e.g., Wu, Y., et al., BioEssays 13(4):163 (1991),
- 2) Human Antithrombin III (ATIII) (potentially useful in the prevention of thrombosis and pulmonary embolism),
- 3) Human Serum Albumin (Geisow, M.J. et al. (1977) Biochem. J. 163:477-484; HSA is used to expand blood volume and raise low blood protein levels in cases of shock, trauma, and post-surgical recovery. HSA is often administered in emergency situations to stabilize blood pressure).
- 4) Subtilisin BPN' is an important industrial enzyme, particularly for use as a detergent enzyme.

5.7.5 PRACTICAL METHOD FOR THE COMMERCIAL PRODUCTION OF TRANSGENIC PLANTS

Translating genetic engineering theory into practice, however, and then furthermore into a commercially practical reality, requires ingenuity. Gene transplantation in plants has already been accomplished at this writing--and examples are cited below--but heretofore no practical method for the commercial production of transgenic plants has been perfected. Genetic engineering of plants may involve any

and/or all of the following steps: tissue culture propagation, gene transplantation (eg., with Agrobacterium and T-DNA), the binary system (using binary vectors)

A general reference is Buchanan, B.B. et al., Biochemistry and Molecular Biology of Plants, ASPP Publications, 2001. Exemplary publications and patents which disclose transgenic plants and various techniques therefor are summarized below. Pellegrineschi, A., et al., Bio/Technology, Vol. 12 (January, 1994) discloses transformation of root cultures by inoculating stem and leaf fragments with Agrobacterium rhizogenes. An important plasmid in this species of Agrobacterium is the root-inducing plasmid which can be used to transfer to the plant genome the genes necessary for improved root growth in culture. The use of sterilized petioles as the source of explant material for plant transformation and culture is disclosed. U.S. Patent No. 5,276,268 to Strauch et al., entitled "11 Phosphinothricin-Resistance Gene, and Its Use," is directed to the transfer of phosphinothricin-resistance gene into plants using Agrobacterium species. A modification of the binary vector method is discussed, and the phosphinothricin-resistance gene nucleic acid sequences are provided. U.S. Patent No. 5,283,184; U.S. Patent No. 5,286,635.

5.7.6 METHOD OF IDENTIFYING AND CHARACTERIZING THE ROLE OF INDIVIDUAL PLANT GENES IN QUANTITATIVE TRAIT EXPRESSION

One area in which biotechnology may have a significant impact on plant improvement is in the development of new methods to identify and characterize the role of individual plant genes in quantitative trait expression. Following the development of a new class of plant molecular markers based on restriction fragment length polymorphisms, termed "RFLPs", (Helentjaris et al., Plant Mol. Bio. 5:109-118 (1985)) ("Helentjaris et al. I"), the processes to identify such loci and discriminate gene effects have been invented and are described and claimed herein. This and all other publications noted herein are hereby incorporated by reference. This will undoubtedly benefit plant improvement, not only within the context of conventional breeding approaches, but also by providing a means for identifying appropriate loci for future cloning and direct gene transfer efforts.

5.7.7. FUSION GENES

The present invention is directed to a DNA construct formed from a fusion gene which includes a trait DNA molecule and a silencer DNA molecule. In an alternative embodiment of the present invention, the DNA construct can be a fusion gene comprising a plurality of trait DNA molecules at least some of which having a length that is insufficient to impart that trait to plants transformed with that trait DNA molecule. However, the plurality of trait DNA molecules collectively have a length sufficient to impart their traits to plants transformed with the DNA construct and to effect post-transcriptional silencing of the fusion gene. Expression systems, host cells, plants, and plant seeds containing this embodiment of the DNA construct are disclosed.

The present invention also provides a recombinant chimeric DNA molecule comprising a plurality of DNA sequences each of which comprises a promoter operably linked to a DNA sequence which encodes a virus-associated protein, such as a coat protein (cp), a protease, or a replicase, wherein said DNA sequences are expressed in virus-susceptible plant cells transformed with said recombinant DNA molecule to impart resistance to infection by each of said viruses. Preferably, the DNA sequences are linked in tandem, i.e., exist in head to tail orientation relative to one another. Also, preferably substantially equal levels of resistance to infection by each of said viruses occurs in plant cells transformed with said plurality of DNA sequences.

Preferably, each DNA sequence is also linked to a 3' non-translated DNA sequence which functions in plant cells to cause the termination of transcription and the addition of polyadenylated ribonucleotides to the 3' end of the transcribed mRNA sequences. Preferably, the virus is a plant-associated virus, such as a potyvirus.

Thus, the present DNA molecule can be employed as a chimeric recombinant "expression construct," or "expression cassette" to prepare transgenic plants that exhibit increased resistance to infection by at least two plant viruses, such as potyviruses. The present cassettes also preferably comprise at least one selectable marker gene or reporter

gene which is stably integrated into the genome of the transformed plant cells in association with the viral genes. The selectable marker and/or reporter genes facilitate identification of transformed plant cells and plants. Preferably, the virus gene array is flanked by two or more selectable marker genes, reporter genes or a combination thereof.

Another aspect of the present invention is a method of preparing a virus-resistant plant, such as a dicot, comprising:

(a) transforming plant cells with a chimeric recombinant DNA molecule comprising a plurality of DNA sequences, each comprising a promoter functional in said plant cells, operably linked to a DNA sequence, which encodes a protein associated with a virus which is capable of infecting said plant; (b) regenerating said plant cells to provide a differentiated plant; and (c) identifying a transformed plant which expresses the DNA sequences so as to render the plant resistant to infection by said viruses, preferably at substantially equal levels of resistance to infection by each virus.

Yet another object of the present invention is to provide a method for providing resistance to infection by viruses in a susceptible Cucurbitaceae plant which comprises:

(a) transforming Cucurbitaceae plant cells with a DNA molecule encoding a plurality of proteins from viruses which are capable of infecting said Cucurbitaceae plant; (b) regenerating said plant cells to provide a differentiated plant; and (c) selecting a transformed Cucurbitaceae which expresses the virus proteins at levels sufficient to render the plant resistant to infection by said viruses.

It is a further object of the present invention to provide multi-virus resistant transformed plant which contains stably-integrated DNA sequences encoding virus proteins.

5.7.8 CONTROLLING GENE EXPRESSION WITH EXTERNAL STIMULUS

The present invention involves, in one embodiment, the creation of a transgenic plant that contains a gene whose expression can be controlled by application of an external stimulus. This system achieves a positive control of gene expression by an external stimulus, without the need for continued application of the external stimulus to maintain gene expression. The present invention also involves, in a second embodiment, the creation of transgenic parental plants that are hybridized to produce a progeny plant expressing a gene not expressed in either parent. By controlling the expression of genes that affect the plant phenotype, it is possible to grow plants under one set of conditions or in one environment where one phenotype is advantageous, then either move the plant or plant its seed under another set of conditions or in another environment where a different phenotype is advantageous. This technique has particular utility in agricultural and horticultural applications.

In accordance with one embodiment of the invention, a series of sequences is introduced into a plant that includes a transiently-active promoter linked to a structural gene, the promoter and structural gene being separated by a blocking sequence that is in turn bounded on either side by specific excision sequences, a repressible promoter operably linked to a gene encoding a site-specific recombinase capable of recognizing the specific excision sequences, and a gene encoding a repressor specific for the repressible promoter whose function is sensitive to an external stimulus. Without application of the external stimulus, the structural gene is not expressed. Upon application of the stimulus, repressor function is inhibited, the recombinase is expressed and effects the removal of the blocking sequence at the specific excision sequences, thereby directly linking the structural gene and the transiently-active promoter.

In a modification of this embodiment, the sequences encoding the recombinase can be introduced separately into the plant via a viral vector.

In an alternative embodiment, no repressor gene or repressible promotor is used. Instead, the recombinase gene is linked to a germination-specific promotor and introduced into a separate plant from the other sequences. The plant containing the transiently-active promotor, blocking sequence, and structural gene is then hybridized with the plant containing the recombinase gene, producing progeny that contain all of the sequences. When the second transiently-active promotor becomes active, the

recombinase removes the blocking sequence in the progeny, allowing expression of the structural gene in the progeny, whereas it was not expressed in either parent.

In still another embodiment, the recombinase gene is simply linked to an inducible promoter. Exposure of the plant to the induce specific for the inducible promoter leads to the expression of the recombinase gene and the excision of the blocking sequence.

In all of these embodiments, the structural gene is expressed when the transiently-active promoter becomes active in the normal course of growth and development, and will continue to be expressed so long as the transiently-active promoter is active, without the necessity of continuous external stimulation. This system is particularly useful for developing seed, where a particular trait is only desired during the first generation of plants grown from that seed, or a trait is desired only in subsequent generations.

5.7.9. PREPARING PLANTS WHICH ARE RESISTANT TO MULTIPLE VIRUSES

It is still a further object of the present invention to provide virus resistant transformed plant cells which contain a plurality of viral genes, i.e., 2-7 or more genes, which are expressed as virus proteins, such as coat proteins, proteases and/or replicases, from the same virus strain, from different virus strains as from different members of the virus group, such as the potyvirus group. Representative viruses from which these DNA sequences can be isolated include, but are not limited to, potato virus X (PVX), potyviruses such as potato virus Y (PVY), cucomovirus (CMV), tobacco vein mottling virus, watermelon mosaic virus (WMV), zucchini yellow mosaic virus (ZYMV), bean common mosaic virus, bean yellow mosaic virus, soybean mosaic virus, peanut mottle virus, beet mosaic virus, wheat streak mosaic virus, maize dwarf mosaic virus, sorghum mosaic virus, sugarcane mosaic virus, johnsongrass mosaic virus, plum pox virus, tobacco etch virus, sweet potato feathery mottle virus, yam mosaic virus, and papaya ringspot virus (PRV), cucomoviruses, including CMA and comovirus.

5.7.3.4.1 Using short fragments of viral genes to impart resistance

Rather than attempting to incorporate full length viral genes in a plant, the present invention uses short fragments of such genes to impart resistance to the plant against a plurality of viral pathogens. These short fragments, which each by themselves have insufficient length to impart such resistance, are more easily and cost effectively produced than full length genes. There is no need to include in the plant separate promoters for each of the fragments; only a single promoter is required. Moreover, such viral gene fragments can preferably be incorporated in a single expression system to produce transgenic plants with a single transformation event.

5.7.10 IMPARTING OTHER TRAITS TO PLANTS

In addition to conferring on plants resistance to multiple viral diseases, the present invention can be utilized to impart other traits to plants. It is often desirable to incorporate a number of traits to a transgenic plant besides disease resistance. For example, color, enzyme production, etc. may be desirable traits to confer on a plant. However, transforming plants with a plurality of such traits encounter the same difficulties discussed above with respect to disease resistance. The present invention may be likewise useful in alleviating these problems with respect to traits other than disease resistance.

Thus, the present invention provides a genetic engineering methodology by which multiple traits can be manipulated and tracked as a single gene insert, i.e., as a construct which acts as a single gene which segregates as a single Mendelian locus. Although the invention is exemplified via virus resistance genes, in practice, any combination of genes could be linked. Therefore one could track a block of genes that provide traits such as disease resistance, plus enhanced herbicide resistance, plus extended shelf life, and the

like, by simply tracking the linked selectable marker or reporter gene which has been incorporated into the transformation vector.

It was also discovered that when multiple tandem genes are inserted, they preferably all exhibit substantially the same degrees of efficacy, and more preferably substantially equal degrees of efficacy, wherein the term "substantial" as it relates to viral resistance is defined with reference to the assays described in the examples hereinbelow. For example, if one examines numerous transgenic lines containing an intact ZYMV and WMV-2 coat protein insert, one finds that if a line is immune to infection by ZYMV it is also immune to infection by WMV-2. Similarly, if a line exhibits a delay in symptom development to ZYMV it will also exhibit a delay in symptom development to WMV2. Finally, if a line is susceptible to ZYMV it will be susceptible to WMV-2. This phenomenon is unexpected. If there were not a correlation between the efficacy of each gene in these multiple gene constructs this approach as a tool in plant breeding would probably be prohibitively difficult to use. Even with single gene constructs, one must test numerous transgenic plant lines to find one that displays the appropriate level of efficacy. The probability of finding a line with useful levels of expression can range from 10-50% (depending on the species involved).

If the efficacy of individual genes in a Ti plasmid containing multiple genes were independent, the probability of finding a transgenic line that was resistant to each targeted virus would decrease dramatically. For example, in a species in which there is a 10% probability of identifying a line with resistance using a single gene insert, is transformed with a triple-gene construct CZW and each gene display an independent levels of efficacy, the probability of finding a line with resistance to CMV, ZYMV and WMV-2 would be $0.1 \times 0.1 \times 0.1 = 0.001$ or 0.1 %. However, since the efficacy of multivalent genes is not independent of each other the probability of finding a line with resistance to CMV, ZYMV and WMV-2 is still 10% rather than 0.1 %. Obviously this advantage becomes more pronounced as constructs containing four or more genes are used.

5.7.10.1. PRODUCTION OF A MATURE HETEROLOGOUS PROTEIN IN TRANSFORMED MONOCOT PLANT CELLS

In one aspect, the invention includes a method of producing, in monocot plant cells, a mature heterologous protein of interest. The method includes obtaining monocot cells transformed with a chimeric gene having (i) a monocot transcriptional regulatory region, inducible by addition or removal of a small molecule, or during seed maturation, (ii) a first DNA sequence encoding the heterologous protein, and (iii) a second DNA sequence encoding a signal peptide. The second DNA sequence is operably linked to the transcriptional regulatory region and to the first DNA sequence. The first DNA sequence is in translation-frame with the second DNA sequence, and the two sequences encode a fusion protein.

5.7.10.2 Inducing the transcriptional regulatory region

In other embodiments of the method, the transcriptional regulatory region may be a promoter derived from a rice or barley α -amylase gene, including RAmy1A, RAmy1B, RAmy2A, RAmy3A, RAmy3B, RAmy3C, RAmy3D, RAmy3E, pM/C, gKAmy4I, gKAmy55, Amy32b, or HV18. The chimeric gene may further include, between the transcriptional regulatory region and the fusion protein coding sequence, the 5' untranslated region (5' UTR) of an inducible monocot gene such as one of the rice or barley α -amylase genes described above. One preferred 5' UTR is that from the RAmy1A gene, which is effective to enhance the stability of the gene transcript. The chimeric gene may further include, downstream of the coding sequence, the 3' untranslated region (3' UTR) from an inducible monocot gene, such as one of the rice or barley α -amylase genes mentioned above. One preferred 3'UTR is from the RAmy1A gene.

Where the method is employed in protein production in a monocot cell culture, preferred promoters are the RAmy3D and RAmy3E gene promoters, which are upregulated by sugar depletion in cell culture. Where the gene is employed in protein production in germinating seeds, a preferred promoter is the RAmyA gene promoter, which is upregulated by gibberellic acid during seed germination. Where gene is upregulated during seed maturation, a preferred promoter is the barley endosperm-specific BI-hordein promoter.

5.7.11 IDENTIFYING SUCH LOCI AND DISCRIMINATING GENE EFFECTS OF RESTRICTION FRAGMENT LENGTH POLYMORPHISMS

The development of new methods to identify and characterize the role of individual plant genes in quantitative trait expression has significant impact on plant improvement. Following the development of a new class of plant molecular markers based on restriction fragment length polymorphisms, termed "RFLPs", (Helentjaris et al., Plant Mol. Bio. 5:109-118 (1985)) ("Helentjaris et al. I"), the processes to identify such loci and discriminate gene effects have been invented and are described and claimed herein. Additional RFLP reference is (Roberts, Nuc. Acids Res. 10:117-144 (1982)). The utility of isozyme markers or morphological markers in studies is frequently limited by a lack of informativeness in lines of interest or by an insufficient availability or chromosomal distribution of the loci. Over 300 RFLPs covering all ten maize chromosomes have been characterized (Helentjaris et al., Trends in Genetics. 3:217-221 (1987)). Various plant genetic linkage maps based on RFLP markers have been constructed (Helentjaris et al., Theor. Appl. Genet. 72:761-769 (1986); Brassica Fiddore et al., Theor. Appl. Genetics. 75:833-840 (1988); Slocum et al., In "Genetic Maps" (S. J. O'Brien, ed.), 5th Edition, Cold Spring Harbor Press, N.Y. (1990); Wright et al., MNL 61:89-90 (1987); Helentjaris et al., Weber and Helentjaris, Genetics 121:583-590 (1989).

5.7.12 Isozyme Variation in Plant Breeding

The use of isozyme variation in plant breeding is, like RFLP technology, one of indirect selection. (Tanksley and Orton, Isozymes in Plant Genetics and Breeding 1B (Elsevier, N.Y. 1983; Vallejos and Tanksley, Theor. Appl. Genet. 66:241-247 (1983); Stuber et al., Crop. Sci. 22:737-740 (1982)).

Maize is perhaps the best characterized plant system in terms of isozymes and yet only about two dozen isozyme loci have been located and it is rare for more than a dozen.

PARTICULAR DEFINITIONS

The terms below have the following meaning, unless indicated otherwise in the specification.

As used in this specification, a "transiently-active promoter" is any promoter that is active either during a particular phase of plant development or under particular environmental conditions, and is essentially inactive at other times.

A "plant active promoter" is any promoter that is active in cells of a plant of interest. Plant-active promoters can be of viral, bacterial, fungal, animal or plant origin.

A gene that results in an altered plant phenotype is any gene whose expression leads to the plant exhibiting a trait or traits that would distinguish it from a plant of the same species not expressing the gene. Examples of such altered phenotypes include a different growth habit, altered flower or fruit color or quality, premature or late flowering, increased or decreased yield, sterility, mortality, disease susceptibility, altered production of secondary metabolites, or an altered crop quality such as taste or appearance.

A gene and a promoter are to be considered to be operably linked if they are on the same strand of DNA, in the same orientation, and are located relative to one another

such that the promoter directs transcription of the gene (i.e. in cis). The presence of intervening DNA sequences between the promoter and the gene does not preclude an operable relationship.

A "blocking sequence" is a DNA sequence of any length that blocks a promoter from effecting expression of a targeted gene.

A "specific excision sequence" is a DNA sequence that is recognized by a site-specific recombinase.

A "recombinase" is an enzyme that recognizes a specific excision sequence or set of specific excision sequences and effects the removal of, or otherwise alters, DNA between specific excision sequences.

A "repressor element" is a gene product that acts to prevent expression of an otherwise expressible gene. A repressor element can comprise protein, RNA or DNA.

A "repressible promoter" is a promoter that is affected by a repressor element, such that transcription of the gene linked to the repressible promoter is prevented.

"Expression" means transcription or transcription followed by translation of a particular DNA molecule.

As used herein, with respect to a DNA sequence or "gene", the term "isolated" is defined to mean that the sequence is either extracted from its context in the viral genome by chemical means and purified and/or modified to the extent that it can be introduced into the present vectors in the appropriate orientation, i.e., sense or antisense.

"Cell culture" refers to cells and cell clusters, typically callus cells, growing on or suspended in a suitable growth medium.

"Germination" refers to the breaking of dormancy in a seed and the resumption of metabolic activity in the seed, including the production of enzymes effective to break down starches in the seed endosperm.

"Inducible" means a promoter that is upregulated by the presence or absence of a small molecules. It includes both indirect and direct inducement.

"Inducible during germination" refers to promoters which are substantially silent but not totally silent prior to germination but are turned on substantially (greater than 25%) during germination and development in the seed. Examples of promoters that are inducible during germination are presented below.

"Small molecules", in the context of promoter induction, are typically small organic or bioorganic molecules less than about 1 kDa. Examples of such small molecules include sugars, sugar-derivatives (including phosphate derivatives), and plant hormones (such as, gibberellic or abscisic acid).

"Specifically regulatable" refers to the ability of a small molecule to preferentially affect transcription from one promoter or group of promoters (e.g., the α-amylase gene family), as opposed to non-specific effects, such as, enhancement or reduction of global transcription within a cell by a small molecule.

"Seed maturation" or "grain development" refers to the period starting with fertilization in which metabolizable reserves, e.g., sugars, oligosaccharides, starch, phenolics, amino acids, and proteins, are deposited, with and without vacuole targeting, to various tissues in the seed (grain), e.g., endosperm, testa, aleurone layer, and scutellar epithelium, leading to grain enlargement, grain filling, and ending with grain desiccation.

"Inducible during seed maturation" refers to promoters which are turned on substantially (greater than 25%) during seed maturation.

"Heterologous" is defined to mean not identical, e.g. different in nucleotide and/or amino acid sequence, phenotype or an independent isolate.

"Heterologous DNA" or "foreign DNA" refers to DNA which has been introduced into plant cells from another source, or which is from a plant source, including the same plant source, but which is under the control of a promoter or terminator that does not normally regulate expression of the heterologous DNA.

"Heterologous protein" is a protein, including a polypeptide, encoded by a heterologous DNA. A "transcription regulatory region" or "promoter" refers to nucleic acid sequences that influence and/or promote initiation of transcription. Promoters are typically considered to include regulatory regions, such as enhancer or inducer elements.

"Chimeric" is defined to mean the linkage of two or more DNA sequences which are derived from different sources, strains or species, i.e., from bacteria and plants, or that two or more DNA sequences from the same species are linked in a way that does not occur in the native genome. Thus, the DNA sequences useful in the present invention may be naturally occurring, semi-synthetic or entirely synthetic. The DNA sequence may be linear or circular, Le, may be located on an intact or linearized plasmid, such as the binary plasmids described below.

A "chimeric gene," in the context of the present invention, typically comprises a promoter sequence operably linked to DNA sequence that encodes a heterologous gene product, e.g., a selectable marker gene or a fusion protein gene. A chimeric gene may also contain further transcription regulatory elements, such as transcription termination signals, as well as translation regulatory signals, such as, termination codons.

"Operably linked" refers to components of a chimeric gene or an expression cassette that function as a unit to express a heterologous protein. For example, a promoter operably linked to a heterologous DNA, which encodes a protein, promotes the production of functional mRNA corresponding to the heterologous DNA.

A "product" encoded by a DNA molecule includes, for example, RNA molecules and polypeptides.

"Removal" in the context of a metabolite includes both physical removal as by washing and the depletion of the metabolite through the absorption and metabolizing of the metabolite by the cells.

"Substantially isolated" is used in several contexts and typically refers to the at least partial purification of a protein or polypeptide away from unrelated or contaminating components.

Methods and procedures for the isolation or purification of proteins or polypeptides are known in the art.

"Stably transformed" as used herein refers to a cereal cell or plant that has foreign nucleic acid stably integrated into its genome which is transmitted through multiple generations.

" α_1 -antitrypsin" or "AAT" refers to the protease inhibitor which has an amino acid sequence substantially identical or homologous to AAT protein.

"Antithrombin III" or "ATIII" refers to the heparin-activated inhibitor of thrombin and factor Xa, and which has an amino acid sequence substantially identical or homologous to ATIII protein.

"Human serum albumin" or "HSA" refers to a protein which has an amino acid sequence substantially identical or homologous to the mature HSA protein.

"Subtilisin" or "subtilisin BPN'" or "BPN'" refers to the protease enzyme produced naturally by *B. amyloliquefaciens*.

"proBPN'" refers to a form of BPN' having an approximately 78 amino-acid "pro" moiety that functions as a chaperon polypeptide to assist in folding and activation of the BPN'.

"Codon optimization" refers to changes in the coding sequence of a gene to replace native codons with those corresponding to optimal codons in the host plant.

A DNA sequence is "derived from" a gene, such as a rice or barley α -amylase gene, if it corresponds in sequence to a segment or region of that gene. Segments of genes

which may be derived from a gene include the promoter region, the 5' untranslated region, and the 3' untranslated region of the gene.

5.7.13 GENERAL APPROACH

Generally, the nomenclature and laboratory procedures with respect to standard recombinant DNA technology can be found in Sambrook, et al., MOLECULAR - CLONING - A LABORATORY MANUAL, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 1989 and in S.B. Gelvin and R.A. Schilperoort, PLANT MOLECULAR BIOLOGY, 1988. Other general references are provided throughout this document. The procedures therein are known in the art and are provided for the convenience of the reader.

Most of the recombinant DNA methods employed in practicing the present invention are standard procedures, well known to those skilled in the art, and described in detail in, or example, European Patent Application Publication Number 223,452, published November 29, 1986, which is incorporated herein by reference. Enzymes are obtained from commercial sources and are used according to the vendor's recommendations or other variations known in the art. General references containing such standard techniques include the following: R. Wu, ed. (1979) Methods Enzymology, Vol. 68; J.H. Miller (1972) Experiments in Molecular Genetics; J. Sambrook et al. (1989) Molecular Cloning: A Laboratory Manual 2nd Ed.; D.M. Glover, ed. (1985) DNA Cloning Vol. II; H.G. Polites and K.R. Marotti (1987) "A step-wise protocol for cDNA synthesis," Biotechniques 4; 514-520; S.B. Gelvin and R.A. Schilperoort, eds. Introduction, Expression, and Analysis of Gene Products in Plants, all of which are incorporated by reference.

The recombinase/excision sequence system can be any one that selectively removes DNA in a plant genome. The excision sequences are preferably unique in the plant, so that unintended cleavage of the plant genome does not occur. Several examples of such systems are discussed in Sauer, U.S. Pat. No. 4,959,317 and in Sadowski (1993).

A preferred system is the bacteriophage CRE/LOX system, wherein the CRE protein performs site-specific recombination of DNA at LOX sites. Other systems include the resolvases (Hall, 1993), FLP (Pan, et al., 1993), SSV1 encoded integrase (Muskhelishvili, et al., 1993), and the maize Ac/Ds transposon system (Shen and Hohn, 1992).

5.7.13.1 CONTROL OF PLANT GENE EXPRESSION

5.7.13.1.1 Using a Transiently-active Promoter

This invention relates to a method of creating transgenic plants wherein the expression of certain plant traits is ultimately under external control. In one embodiment the control is achieved through application of an external stimulus; in another embodiment it is achieved through hybridization, in still another embodiment it is achieved by direct introduction of a recombinase or recombinase gene into a plant. The transgenic plants of the present invention are prepared by introducing into their genome a series of functionally interrelated DNA sequences, containing the following basic elements: a plant-active promoter that is active at a particular stage in plant development or under particular environmental conditions ("transiently- active promoter"), a gene whose expression results in an altered plant phenotype which is linked to the transiently-active promoter through a blocking sequence separating the transiently-active promoter and the gene, unique specific excision sequences flanking the blocking sequence, wherein the specific excision sequences are recognizable by a site- specific recombinase, a gene encoding the site-specific recombinase, an alternative repressible promoter linked to the recombinase gene, and an alternative gene that encodes the repressor specific for the repressible promoter, the action of the repressor being responsive to an applied or exogenous stimulus. While these elements may be arranged in any order that achieves the interactions described below, in one embodiment they are advantageously arranged as follows: a first DNA sequence contains the transiently-active promotor, a first specific excision sequence, the blocking sequence, a second specific excision sequence, and the gene whose expression results in an altered plant phenotype; a second DNA sequence

contains the repressible promoter operably linked to the recombinase gene, and optionally an enhancer; and a third DNA sequence containing the gene encoding the repressor specific for the repressible promoter, itself linked to a promoter functional and constitutive in plants. The third DNA sequence can conveniently act as the blocking sequence located in the first DNA sequence, but can also occur separately without altering the function of the system. This embodiment can be modified such that the recombinase sequence is introduced separately via a viral vector. In an alternative embodiment, an advantageous arrangement is as follows: a first plant containing a DNA sequence comprising the transiently-active promotor, a first specific excision signal sequence, the blocking sequence, a second specific excision signal sequence, and the gene whose expression results in an altered plant phenotype; a second plant containing a DNA sequence comprising a constitutive plant-active promotor operably linked to the recombinase gene (the two plants being hybridized to produce progeny that contain all of the above sequences).

When a plant contains the basic elements of either embodiment, the gene whose expression results in an altered plant phenotype is not active, as it is separated from its promoter by the blocking sequence. In the first embodiment, absent the external stimulus, the repressor is active and represses the promoter that controls expression of the recombinase; in the alternative embodiment the recombinase is not present in the same plant as the first DNA sequence. Such a plant will not display the altered phenotype, and will produce seed that would give rise to plants that also do not display the altered phenotype. When the stimulus to which the repressor is sensitive is applied to this seed or this plant, the repressor no longer functions, permitting the expression of the site-specific recombinase, or alternatively, when the recombinase is introduced via hybridization it is expressed during germination of the seed, either of which effects the removal of the blocking sequence between the specific excision signal sequences. Upon removal of the blocking sequence, the transiently-active promoter becomes directly linked to the gene whose expression results in an altered plant phenotype. A plant grown from either treated or hybrid seed, or a treated plant, will still not exhibit the altered phenotype, until the transiently-active promoter becomes active during the plant's

development, after which the gene to which it is linked is expressed, and the plant will exhibit an altered phenotype.

5.7.13.2 Transgenic Plants that Produce Seeds that Cannot Germinate

In a preferred embodiment, the present invention involves a transgenic plant or seed which, upon treatment with an external stimulus produces plants that produce seed that cannot germinate (but that is unaltered in other respects). If the transiently-active promoter is one that is active only in late embryogenesis, the gene to which it is linked will be expressed only in the last stages of seed development or maturation. If the gene linked to this promoter is a lethal gene, it will render the seed produced by the plants incapable of germination. In the initially-transformed plant cells, this lethal gene is not expressed, not only because the promoter is intrinsically inactive, but because there is a blocking sequence separating the lethal gene from its promoter. Also within the genome of these cells are the genes for the recombinase, linked to a repressible promoter, and the gene coding for the repressor. The repressor is expressed constitutively and represses the expression of the recombinase. These plant cells can be regenerated into a whole plant and allowed to produce seed. The mature seed is exposed to a stimulus, such as a chemical agent, that inhibits the function of the repressor. Upon inhibition of the repressor, the promoter driving the recombinase gene is depressed and the recombinase gene is expressed. The resulting recombinase recognizes the specific excision sequences flanking the blocking sequence, and effects the removal of the blocking sequence. The late embryogenesis promoter and the lethal gene are then directly linked. The lethal gene is not expressed, however, because the promoter is not active at this time in the plant's life cycle. This seed can be planted, and grown to produce a desired crop of plants. As the crop matures and produces a second generation of seed, the late embryogenesis promoter becomes active, the lethal gene is expressed in the maturing second generation seed, which is rendered incapable of germination. In this way, accidental reseeding, escape of the crop plant to areas outside the area of cultivation, or germination of stored seed can be avoided.

5.7.13.1.3 Trangenic Plants Hybridized to Display Phenotype Not Seen in Either Parent

In an alternative preferred embodiment, the present invention involves a pair of transgenic plants that are hybridized to produce progeny that display a phenotype not seen in either parent. In this alternative embodiment a transiently-active promotor that is active only in late embryogenesis can be linked to a lethal gene, with an intervening blocking sequence bounded by the specific excision sequences. These genetic sequences can be introduced into plant cells to produce one transgenic parent plant. The recombinase gene is linked to a germination-specific promotor and introduced into separate plant cells to produce a second transgenic parent plant. Both of these plants can produce viable seed if pollinated. If the first and second transgenic parent plants are hybridized, the progeny will contain both the blocked lethal gene and the recombinase gene. The recombinase is expressed upon germination of the seed and effects the removal of the blocking sequence, as in the first embodiment, thereby directly linking the lethal gene and the transiently-active promotor. As in the first embodiment, this promotor becomes active during maturation of the second generation seed, resulting in seed that is incapable of germination. Ideally, the first parent employs a male-sterility gene as the blocking sequence, and includes an herbicide resistance gene. In this way, self-pollination of the first transgenic parent plant is avoided, and self-pollinated second transgenic parent plants can be eliminated by application of the herbicide. In the hybrid progeny, the male-sterility gene is removed by the recombinase, resulting in hybrid progeny capable of self-pollination.

5.7.13.1.4. Linking the Recombinase Gene to an Inducible Promoter

In another embodiment, the recombinase gene is linked to an inducible promoter. Examples of such promoters include the copper, controllable gene expression system (Mett et al., 1993) and the steroid- inducible gene system (Schena et al., 1991). Exposure of the transgenic plant to the inducer specific for the inducible promoter leads to expression of the recombinase gene and the excision of the blocking sequence. The gene that results in an altered plant phenotype is then expressed when the transiently active promoter becomes active.

The gene or genes linked to the plant development promoter can be any gene or genes whose expression results in a desired detectable phenotype. In one method, Gene(s) can be Linked to the Plant Development Promoter to control expression developmentally.

In those embodiments employing a repressible promoter system, the gene encoding the repressor is responsive to an outside stimulus, or encodes a repressor element that is itself responsive to an outside stimulus, so that repressor function can be controlled by the outside stimulus. The stimulus is preferably one to which the plant is not normally exposed, such as a particular chemical, temperature shock, or osmotic shock. A preferred system is the Tn10 tet repressor system, which is responsive to tetracycline. Gatz and Quail (1988); Gatz, et al. (1992). Examples of other repressible promoter systems are described by Lanzer and Bujard (1988) and Ptashne, et al.

5.7.14. CONFERRING VIRAL RESISTANCE TO THE PLANT

To practice the present invention, a viral gene must be isolated from the viral genome and inserted into a vector containing the genetic regulatory sequences necessary to express the inserted gene. Accordingly, a vector must be constructed to provide the regulatory sequences such that they will be functional upon inserting a desired gene. When the expression vector/insert construct is assembled, it is used to transform plant cells which are then used to regenerate plants. These transgenic plants carry the viral gene in the expression vector/insert construct. The gene is expressed in the plant and increased resistance to viral infection is conferred thereby.

5.7.14.1. Table of Selected Literature References to Methods of Isolating, cloning and Expressing Viral Genes

The nucleotide sequences encoding the coat protein genes and nuclear inclusion genes of a number of viruses have been determined and the genes have been inserted into expression vectors. The expression vectors contain the necessary genetic regulatory sequences for expression of an inserted gene. The coat protein gene is inserted such that those regulatory sequences are functional and the genes can be expressed when

incorporated into a plant genome. Selected literature references to methods of isolating, cloning and expressing viral genes are listed on Table 3, below.

TABLE 3.

Cloned Genes From RNA Viruses

Viral Gene	Reference
Papaya ringspot cp	M.M. Fitch et al., Bio/Technology, 10, 1466(1992)
Potato virus X cp	K. Ling et al., Bio/Technology, 2, 752 (1991); A. Hoekema et al., Bio/Technology, 7, 273 (1989)
Watermelon Mosaic Virus II cp	H. Quemada et al., J. Gen. Virol., 71, 1451 (1990); S. Namba et al., Phytopathology, 82, 940 (1992)
Zucchini yellow Mosaic Virus cp	S. Namba et al., Phytopathology, 82, 940 (1992)
Tobacco Mosaic Virus cp	R.S. Nelson et al., Bio/Technology, 6, 403 (1988); P. Powell Abel et al., Science, 232, 738 (1986)
Alfalfa Mosaic Virus cp	Loesch-Fries et al., EMBO J., 6, 1845 (1987); N.E. Turner et al., EMBO J., 6, 1181 (1987)
Soybean Mosaic Virus cp	D.M. Stark et al., Biotechnology, 7, 1257 (1989)
Cucumber Mosaic Virus strain C cp	H.Q. Quemada et al., Molec. Plant Pathol., 81, 794 (1991)

Cucumber Mosaic Virus strain WL cp	UpJohn Co. (PCT W090/02185)
Tobacco etch virus cp	Allison et al., Virology, 147, 309 (1985)
Tobacco etch virus nuclear inclusion protein	J.C. Carrington et al., J. Virol., 61, 2540 (1987)
Pepper Mottle Virus cp	W.G. Dougherty et al., Virology, 146, 282 (1985)
Potato virus Y cp	D.D. Shukla et al., Virology, 152, 118 (1986)
Potato virus Y nuclear inclusion protein	European Patent Application 578,627
Potato virus X cp	C. Lawson et al., Biotechnology, 8, 127 (1990)
Tobacco streak virus (TSV) cp	C.M. Van Dun et al., Virology, 164, 383 (1988)

5.7.15 DISEASE RESISTANCE

One aspect of the present invention relates to the use of trait DNA molecules which are heterologous to the plant -- e.g., DNA molecules that confer disease resistance to plants transformed with the DNA construct.

5.7.15 .1 Sense/Antisense orientation

The DNA molecule conferring disease resistance can be positioned within the DNA construct in sense orientation. Alternatively, it can have an antisense orientation. Antisense RNA technology involves the production of an RNA molecule that is complementary to the messenger RNA molecule of a target gene; the antisense RNA can potentially block all expression of the targeted gene. In the anti-virus context, plants are made to express an antisense RNA molecule corresponding to a viral RNA (that is, the antisense RNA is an RNA molecule which is complementary to a plus sense RNA

species encoded by an infecting virus). Such plants may show a slightly decreased susceptibility to infection by that virus. Such a complementary RNA molecule is termed antisense RNA.

5.7.16 OTHER TRAITS

The present invention is also used to confer traits other than disease resistance on plants. For example, DNA molecules which impart a plant genetic trait can be used as the DNA trait molecule of the present invention. In this aspect of the present invention, suitable trait DNA molecules encode for desired color, enzyme production, or combinations thereof.

5.7.16.1 GENE SILENCING

The silencer DNA molecule of the present invention can be selected from virtually any nucleic acid which effects gene silencing. This involves the cellular mechanism to degrade mRNA homologous to the transgene mRNA. The silencer DNA molecule can be heterologous to the plant, need not interact with the trait DNA molecule in the plant, and can be positioned 3' to the trait DNA molecule. For example, the silencer DNA molecule can be a viral cDNA molecule, a jellyfish green fluorescence protein encoding DNA molecule, a plant DNA molecule, or combinations thereof.

While not wishing to be bound by theory, by use of the construct of the present invention, it is believed that post-transcriptional gene silencing is achieved. More particularly, the silencer DNA molecule is believed to boost the level of heterologous RNA within the cell above a threshold level. This activates the degradation mechanism by which viral resistance is achieved.

5.7.16.1.1 TRAIT & SILENCER DNA MOLECULES ENCODING RNA MOLECULES - TRANSLATABLE

It is possible for the DNA construct of the present invention to be configured so that the trait and silencer DNA molecules encode RNA molecules which are translatable. As a result, that RNA molecule will be translated at the ribosomes to produce the protein encoded by the DNA construct. Production of proteins in this manner can be increased by joining the cloned gene encoding the DNA construct of interest with synthetic double-stranded oligonucleotides which represent a viral regulatory sequence (i.e., a 5' untranslated sequence). See U.S. Patent No. 4,820,639 to Gehrke and U.S. Patent No. 5,849,527 to Wilson which are hereby incorporated by reference.

5.7.16.1.2 TRAIT & SILENCER DNA MOLECULES ENCODING RNA MOLECULES – NOT TRANSLATABLE

Alternatively, the DNA construct of the present invention can be configured so that the trait and silencer DNA molecules encode mRNA which is not translatable. This is achieved by introducing into the DNA molecule one or more premature stop codons, adding one or more bases (except multiples of 3 bases) to displace the reading frame, removing the translation initiation codon, etc. See U.S. Patent No. 5,583,021 to Dougherty et al., which is hereby incorporated by reference.

5.7.17 RECOMBINANT DNA TECHNOLOGY

The subject DNA construct can be incorporated in cells using conventional recombinant DNA technology. Generally, this involves inserting the DNA construct into an expression system to which the DNA construct is heterologous (i.e. not normally present). The heterologous DNA construct may be inserted into the expression system or

vector in proper sense orientation and correct reading frame. The vector contains the necessary elements for the transcription of the inserted sequences.

5.7.18 INCORPORATION INTO A HOST CELL

Once the DNA construct has been cloned into an expression system, it is ready to be incorporated into a host cell. Such incorporation can be carried out by the various forms of transformation noted above, depending upon the vector/host cell system. Suitable host cells include, but are not limited to, bacteria, virus, plant, is and the like cells.

5.7.18.1 PLANT TRANSFORMATION - PRODUCTION OF MATURE PROTEINS IN PLANTS

Expression vectors for use in the present invention comprise a chimeric gene (or expression cassette), designed for operation in plants, with companion sequences upstream and downstream from the expression cassette. The companion sequences will be of plasmid or viral origin and provide necessary characteristics to the vector to permit the vectors to move DNA from bacteria to the desired plant host. For transformation of plants, the chimeric gene is placed in a suitable expression vector designed for operation in plants. The vector includes suitable elements of plasmid or viral origin that provide necessary characteristics to the vector to permit the vectors to move DNA from bacteria to the desired plant host. Suitable components of the expression vector, including an inducible promoter, coding sequence for a signal peptide, coding sequence for a mature heterologous protein, and suitable termination sequences, are discussed below. One exemplary vector is the p3Dvl.O (p3D(AAT)v1.0) vector described herein.

5.7.18.1.2 Transformation Vector

Vectors containing a chimeric gene of the present invention may also include selectable markers for use in plant cells (such as the nptII kanamycin resistance gene, for selection in kanamycin-containing or the phosphinothricin acetyltransferase gene, for selection in medium containing phosphinothricin (PPT)).

The vectors may also include sequences that allow their selection and propagation in a secondary host, such as sequences containing an origin of replication and a selectable marker such as antibiotic or herbicide resistance genes, e.g., HPH (Hagio et al., Plant Cell Reports 14:329 (1995); van der Elzer, Plant Mol. Biol. 5:299-302 (1985)). Typical secondary hosts include bacteria and yeast. In one embodiment, the secondary host is *Escherichia coli*, the origin of replication is a colEl-type, and the selectable marker is a gene encoding ampicillin resistance. Such sequences are well known in the art and are commercially available as well (e.g., Clontech, Palo Alto, CA; Stratagene, La Jolla, CA).

The vectors of the present invention may also be modified to intermediate plant transformation plasmids that contain a region of homology to an Agrobacterium tumefaciens vector, a T-DNA border region from Agrobacterium tumefaciens, and chimeric genes or expression cassettes (described above). Further, the vectors of the invention may comprise a disarmed plant tumor inducing plasmid of Agrobacterium tumefaciens.

5.7.19 PLANT EXPRESSION VECTOR PRODUCTION OF MATURE PROTEINS IN PLANTS

5.7.19.1 SUITABLE VECTORS

Suitable vectors include, but are not limited to, the following viral vectors such as lambda vector system gt11, gt WES.tB, Charon 4, and plasmid vectors such as pER322, pBR325, pACYC177, pACYC1084, pUC8, pUC9, pUC18, pUC19, pLG339, pR290, pKC37, pKC101, SV 40, pBluescript II SK +/- or KS +/- (see "Stratagene Cloning Systems" Catalog (1993) from Stratagene, La Jolla, Calif, which is hereby incorporated

by reference), pQE, pIH821, pGEX, pET series (see F.W. Studier et. al., "Use of T7 RNA Polymerase to Direct Expression of Cloned Genes," Gene Expression Technolog vol. 185 (1990), which is hereby incorporated by reference), and any derivatives thereof. Recombinant molecules can be introduced into cells via transformation, particularly transduction, conjugation, mobilization, or electroporation. The DNA sequences are cloned into the vector using standard cloning procedures in the art, as described by Sambrook et al., Molecular Cloning: A Laboratory Manual, Cold Springs Laboratory, Cold Springs Harbor, New York (1989), which is hereby incorporated by reference.

5.7.19.2 HOST-VECTOR SYSTEMS

A variety of host-vector systems may be utilized to carry out the present invention. Primarily, the vector system must be compatible with the host cell used. Host-vector systems include but are not limited to the following: bacteria transformed with bacteriophage DNA, plasmid DNA, or cosmid DNA; microorganisms such as yeast containing yeast vectors; mammalian cell systems infected with virus (e.g., vaccinia virus, adenovirus, etc.); insect cell systems infected with virus (e.g., baculovirus); and plant cells infected by bacteria. The expression elements of these vectors vary in their strength and specificities. Depending upon the host- vector system utilized, any one of a number of suitable transcription and, perhaps, translation elements can be used.

5.7.19.3 Signal Sequences for production of mature proteins

In addition to encoding the protein of interest, the chimeric gene encodes a signal sequence (or signal peptide) that allows processing and translocation of the protein, as appropriate. Suitable signal sequences are described in above-referenced PCT application WO 95/14099. The plant signal sequence is placed in frame with a heterologous nucleic

acid encoding a mature protein, forming a construct which encodes a fusion protein having an N-terminal region corresponding to the signal peptide and, immediately adjacent to the C-terminal amino acid of the signal peptide, the N-terminal amino acid of the mature heterologous protein. The expressed fusion protein is subsequently secreted and processed by signal peptidase cleavage precisely at the junction of the signal peptide and the mature protein, to yield the mature heterologous protein.

In another embodiment of the invention, the coding sequence in the fusion protein gene, in at least the coding region for the signal sequence, may be codon-optimized for optimal expression in plant cells, e.g., rice cells, as described below.

5.7.20 PROMOTORS

Transcription dependent upon the presence of a promotor

Transcription of DNA is dependent upon the presence of a promotor which is a DNA sequence that directs the binding of RNA polymerase and thereby promotes mRNA synthesis. The DNA sequences of eucaryotic promoters differ from those of prokaryotic promoters. Furthermore, eucaryotic promoters and accompanying genetic signals may not be recognized in or may not function in a prokaryotic system, and, further, prokaryotic promoters are not recognized and do not function in eucaryotic cells.

The segment of DNA referred to as the promoter is responsible for the regulation of the transcription of DNA into mRNA. A number of promoters which function in plant cells are known in the art and may be employed in the practice of the present invention. These promoters may be obtained from a variety of sources such as plants or plant viruses, and may include but are not limited to promoters isolated from the caulimovirus group such as the cauliflower mosaic virus 35S promoter (CaMV35S), the enhanced cauliflower mosaic virus 35S promoter (enh CaMV35S), the figwort mosaic virus full-length transcript promoter (FMV35S), and the promoter isolated from the chlorophyll alb binding protein. Other useful promoters include promoters which are capable of

expressing the potyvirus proteins in an inducible manner or in a tissue-specific manner in certain cell types in which the infection is known to occur. For example, the inducible promoters from phenylalanine ammonia lyase, chalcone synthase, hydroxyproline rich glycoprotein, extensin, pathogenesis-related proteins (e.g. PR-1a), and wound-inducible protease inhibitor from potato may be useful.

Preferred promoters for use in the present viral gene expression cassettes include the constitutive promoters from CaMV, the Ti genes nopaline synthase (Bevan et al., Nucleic Acids Res. II, 369-385 (1983)) and octopine synthase (Depicker et al., J. Mol. Appl. Genet., 1, 561- 564 (1982)), and the bean storage protein gene phaseolin. The poly(A) addition signals from these genes are also suitable for use in the present cassettes. The particular promoter selected is preferably capable of causing sufficient expression of the DNA coding sequences to which it is operably linked, to result in the production of amounts of the proteins or the RNAs effective to provide viral resistance, but not so much as to be detrimental to the cell in which they are expressed. The promoters selected should be capable of functioning in tissues including but not limited to epidermal, vascular, and mesophyll tissues. The actual choice of the promoter is not critical, as long as it has sufficient transcriptional activity to accomplish the expression of the preselected proteins or antisense RNA, and subsequent conferral of viral resistance to the plants.

Promotors vary in their "strength" (i.e. their ability to promote transcription). For the purposes of expressing a cloned gene, it is desirable to use strong promotors in order to obtain a high level of transcription and, hence, expression of the gene. Depending upon the host cell system utilized, any one of a number of suitable promotors may be used. For instance, when cloning in *E. coli*, its bacteriophages, or plasmids, promotors such as the T7 phage promoter, lac promotor, trp promotor, recA promotor, ribosomal RNA promotor, the P_R and P_L promotors of coliphage lambda and others, including but not limited, to lacUV5, ompF, bla, 1pp, and the like, may be used to direct high levels of transcription of adjacent DNA segments. Additionally, a hybrid trp-lacUV5 (tac) promotor or other *E. coli* promotors produced by recombinant DNA or other synthetic DNA techniques may be used to provide for transcription of the inserted gene.

5.7.20.1 Promoters that transcribe the cereal α -amylase genes and sucrose synthase genes

The transcription regulatory or promoter region is chosen to be regulated in a manner allowing for induction under selected cultivation conditions, e.g., sugar depletion in culture or water uptake followed by gibberellic acid production in germinating seeds. Suitable promoters, and their method of selection are detailed in above-cited PCT application WO 95/14099. Examples of such promoters include those that transcribe the cereal α -amylase genes and sucrose synthase genes, and are repressed or induced by small molecules, like sugars, sugar depletion or phytohormones such as gibberellic acid or abscisic acid. Representative promoters include the promoters from the rice α -amylase RAmy1A, RAmy1B, RAmy2A, RAmy3A, RAmy3B, RAmy3C, RAmy3D, and RAmy3E genes, and from the pM/C, gKAmy4l, gKAmy55, Amy32b, and HV18 barley (α -amylase genes. These promoters are described, for example, in ADVANCES IN PLANT BIOTECHNOLOGY Ryu, D.D.Y., et al, Eds., Elsevier, Amsterdam, 1994, p.37, and references cited therein. Other suitable promoters include the sucrose synthase and sucrose-6-phosphate-synthetase (SPS) promoters from rice and barley.

Other suitable promoters include promoters which are regulated in a manner allowing for induction under seed-maturation conditions. Examples of such promoters include those associated with the following monocot storage proteins: rice glutelins, oryzins, and prolamines, barley hordeins, wheat gliadins and glutelins, maize zeins and glutelins, oat glutelins, and sorghum kafirins, millet pennisetins, and rye secalins.

A preferred promoter for expression in germinating seeds is the rice α -amylase RAmy1A promoter, which is upregulated by gibberellic acid. Preferred promoters for expression in cell culture are the rice α -amylase RAmy3D and RAmy3E promoters which are strongly upregulated by sugar depletion in the culture. These promoters are also active during seed germination. A preferred promoter for expression in maturing seeds is

the barley endosperm-specific BI-hordein promoter (Brandt, A., et al., (1985) Carlsberg Res. Commun. 50:333-345).

The chimeric gene may further include, between the promoter and coding sequences, the 5' untranslated region (5' UTR) of an inducible monocot gene, such as the 5' UTR derived from one of the rice or barley α -amylase genes mentioned above. One preferred 5' UTR is that derived from the RAmylA gene, which is effective to enhance the stability of the gene transcript.

5.7.20.2 Use of inducers

Bacterial host cell strains and expression vectors may be chosen which inhibit the action of the promotor unless specifically induced. In certain operations, the addition of specific inducers is necessary for efficient transcription of the inserted DNA. For example, the lac operon is induced by the addition of lactose or IPTG (isopropylthio-beta-D- galactoside). A variety of other operons, such as trp, pro, etc., are under different controls.

5.7.20.3 TRANSCRIPTION INITIATION SIGNALS

Specific initiation signals are also required for efficient gene transcription in procaryotic cells.

These transcription initiation signals may vary in "strength" as measured by the quantity of gene specific messenger RNA and protein synthesized, respectively. The DNA expression vector, which contains a promotor, may also contain any combination of various "strong" transcription initiation signals.

5.7.20.4 Translation dependent on Shine-Dalgarno sequence (in prokaryotes)

Similarly, translation of mRNA in prokaryotes depends upon the presence of the proper prokaryotic signals which differ from those of eukaryotes. Efficient translation of mRNA in prokaryotes requires a ribosome binding site called the Shine-Dalgarno ("SD") sequence on the mRNA. This sequence is a short nucleotide sequence of mRNA that is located before the start codon, usually AUG, which encodes the amino-terminal methionine of the protein. The SD sequences are complementary to the 3' end of the 16S rRNA (ribosomal RNA) and probably promote binding of mRNA to ribosomes by duplexing with the rRNA to allow correct positioning of the ribosome. For a review on maximizing gene expression, see Roberts and Lauer, *Methods in Enzymology*, 68:473 (1979), which is hereby incorporated by reference.

The non-translated leader sequence can be derived from any suitable source and can be specifically modified to increase the translation of the mRNA. The 5' non-translated region can be obtained from the promoter selected to express the gene, an unrelated promoter, the native leader sequence of the gene or coding region to be expressed, viral RNAs, suitable eukaryotic genes, or a synthetic gene sequence. The present invention is not limited to the constructs presented in the following examples.

5.7.20.5 A4. Codon-Optimized Coding Sequences

In accordance with one aspect of the invention, it has been discovered that a severalfold enhancement of expression level can be achieved in plant cell culture by modifying the native coding sequence of a heterologous gene by containing predominantly or exclusively, highest-frequency codons found in the plant cell host.

The method will be illustrated for expression of a heterologous gene in rice plant cells, it being recognized that the method is generally applicable to any monocot. As a first step, a representative set of known coding gene sequence from rice is assembled. The sequences are then analyzed for codon frequency for each amino acid, and the most

frequent codon is selected for each amino acid. This approach differs from earlier reported codon matching methods, in which more than one frequent codon is selected for at least some of the amino acids. The optimal codons selected in this manner for rice and barley are shown in Table 4.

TABLE 4

Amino Acid	Rice Preferred Codon	Barley Preferred Codon
Ala A	GCC	
Arg R	CGC	
Asn N	AAC	
Asp D	GAC	
Cys C	UGC	
Gln Q	CAG	
Glu E	GAG	
Gly G	GGC	
His H	CAC	
Ile I	AUC	
Leu L	CUC	
Lys K	AAG	
Phe F	UUC	
Pro P	CCG	CCC
Ser S	AGC	UCC
Thr T	ACC	
Tyr Y	UAC	
Val V	GUC	GUG
stop	UAA	UGA

As indicated above, the fusion protein coding sequence in the chimeric gene is constructed such that the final (C-terminal) codon in the signal sequence is immediately followed by the codon for the N-terminal amino acid in the mature form of the heterologous protein.

TABLE 5

N-Glycosylation Sites	Location (Asn) (in mature protein)	Amino Acid Substitution
Asn Asn Ser	61	Thr Asn Ser
Asn Asn Ser	76	Thr Asn Ser
Asn Met Ser	123	Thr Met Ser
Asn Gly Thr	218	Ser Gly Thr'
Asn Trp Thr	240	Thr Trp Thr

'improved thermostability; Bryan, et al., Proteins: Structure, Function, and Genetics 1:326 (1986).

5.7.20.6 TERMINATION SEQUENCE COUPLED TO THE FUSION END

The present invention can also utilize a termination sequence operatively coupled to the fusion gene to end transcription. Suitable transcription termination sequences include the termination region of a 3' non-translated region. This will cause the termination of transcription and the addition of polyadenylated ribonucleotides to the 3' end of the transcribed mRNA sequence. The termination region or 31 non-translated region will be additionally one of convenience. The termination region may be native with the promoter region or may be derived from another source, and preferably includes a terminator and a sequence coding for polyadenylation. Suitable 3' non-translated regions include but are not limited to: (1) the 3' transcribed, non-translated regions

containing the polyadenylated signal of Agrobacterium tumor-inducing (Ti) plasmid genes, such as the nopaline synthase (NOS) gene or the 35S promoter terminator gene, and (2) plant genes like the soybean 7S storage protein genes and the pea small subunit of the ribulose 1,5-bisphosphate carboxylase-oxygenase (ssRUBISCO) E9 gene.

The termination region or 3' non-translated region which is employed is one which will cause the termination of transcription and the addition of polyadenylated ribonucleotides to the 3' end of the transcribed mRNA sequence. The termination region may be native with the promoter region, native with the structural gene, or may be derived from another source, and preferably include a terminator and a sequence coding for polyadenylation. Suitable 3' non-translated regions of the chimeric plant gene include but are not limited to: (1) the 3' transcribed, non-translated regions containing the polyadenylation signal of Ag_jQbacterium tumor-inducing (Ti) plasmid genes, such as the nopaline synthase (NOS) gene, and (2) plant genes like the soybean 7S storage protein genes.

5.7.20.7 Transcription and Translation Terminators for production of mature proteins

The chimeric gene may also include, downstream of the coding sequence, the 3' untranslated region (Y UTR) from an inducible monocot gene, such as one of the rice or barley α-amylase genes mentioned above. One preferred 3' UTR is that derived from the RAmylA gene. This sequence includes non-coding sequence 5' to the polyadenylation site, the polyadenylation site, and the transcription termination sequence. The transcriptional termination region may be selected, particularly for stability of the mRNA to enhance expression. Polyadenylation tails (Alber and Kawasaki, 1982, Mol. and Appl. Genet. 1:419-434) are also commonly added to the expression cassette to optimize high levels of transcription and proper transcription termination, respectively. Polyadenylation sequences include but are not limited to the Agrobacterium octopine synthetase signal

(Gielen, et al., EMBO J. 3:835- 846 (1984) or the nopaline synthase of the same species (Depicker, et al. , Mol. Appl. Genet. 1:561- 573 (1982).

Since the ultimate expression of the heterologous protein will be in a eukaryotic cell (in this case, a member of the grass family), it is desirable to determine whether any portion of the cloned gene contains sequences which will be processed out as introns by the host's splicing machinery. If so, site-directed mutagenesis of the "intron" region may be conducted to prevent losing a portion of the genetic message as a false intron code (Reed and Maniatis, Cell 41:95-105 (1985)).

5.7.20.9 SELECTABLE MARKER GENE

Selectable marker genes may be incorporated into the present expression cassettes and used to select for those cells or plants which have become transformed. The marker gene employed may express resistance to an antibiotic, such as kanamycin, gentamycin, G418, hygromycin, streptomycin, spectinomycin, tetracycline, chloramphenicol, and the like.

Other markers could be employed in addition to or in the alternative, such as, for example, a gene coding for herbicide tolerance such as tolerance to glyphosate, sulfonylurea, phosphinothrinicin, or bromoxynil. Additional means of selection could include resistance to methotrexate, heavy metals, complementation providing prototrophy to an auxotrophic host, and the like.

For example, see Table 1 of PCT WO/91/10725, cited above. The present invention also envisions replacing all of the virus-associated genes with an array of selectable marker genes.

The particular marker employed will be one which will allow for the selection of transformed cells as opposed to those cells which were not transformed. Depending on the number of different host species one or more markers may be employed, where

different conditions of selection would be useful to select the different host, and would be known to those of skill in the art. A screenable marker or "reporter gene" such as the β -glucuronidase gene or luciferase gene may be used in place of, or with, a selectable marker. Cells transformed with this gene may be identified by the production of a blue product on treatment with 5-bromo-4-chloro-3-indoyl- β -D-glucuronide (X-Gluc).

In developing the present expression construct, the various components of the expression construct such as the DNA sequences, linkers, or fragments thereof will normally be inserted into a convenient cloning vector, such as a plasmid or phage, which is capable of replication in a bacterial host, such as *E. coli*. Numerous cloning vectors exist that have been described in the literature. After each cloning, the cloning vector may be isolated and subjected to further manipulation, such as restriction, insertion of new fragments, ligation, deletion, resection, insertion, in vitro mutagenesis, addition of polylinker fragments, and the like, in order to provide a vector which will meet a particular need.

5.7.20.10 TRANSFERRING RECOMBINANT DNA INTO PLANT CELL

5.7.20.10.1 Use of micropipettes or polyethylene glycol

In producing transgenic plants, the DNA construct in a vector described above can be microinjected directly into plant cells by use of micropipettes to transfer mechanically the recombinant DNA. Crossway, Mol. Gen. Genetics, 202:179-85 (1985), which is hereby incorporated by reference. The genetic material may also be transferred into the plant cell using polyethylene glycol. Krens, et al., Nature, 296:72-74 (1982), which is hereby incorporated by reference.

5.7.4.12.2 Particle Bombardment (Biolistic Transformation)

Another approach to transforming plant cells with the DNA construct is particle bombardment (also known as biolistic transformation) of the host cell. This can be accomplished in one of several ways. The first involves propelling inert or biologically active particles at cells. This technique is disclosed in U.S. Patent Nos. 4,945,050, 5,036,006, and 5,100,792, all to Sanford et al., which are hereby incorporated by reference.

5.7.4.12.3 Fusion of protoplasts with other entities

Yet another method of introduction is fusion of protoplasts with other entities,- either minicells, cells, lysosomes or other fusible lipid-surfaced bodies. Fraley, et al., Proc. Natl. Acad. Sci. USA, 79:1859-63 (1982), which is hereby incorporated by reference.

5.7.4.12.4 Electroporation

The DNA molecule may also be introduced into the plant cells by electroporation. Fromm et al., Proc. Natl. Acad. Sci. USA, 82:5824 (1985), which is hereby incorporated by reference. In this technique, plant protoplasts are electroporated in the presence of plasmids containing the expression cassette. Electrical impulses of high field strength reversibly permeabilize biomembranes allowing the introduction of the plasmids. Electroporated plant protoplasts reform the cell wall, divide, and regenerate.

5.7.4.12.5 Infection with Agrobacterium tumefaciens or A. rhizogenes

Another method of introducing the DNA molecule into plant cells is to infect a plant cell with Agrobacterium tumefaciens or A. rhizogenes previously transformed with the gene. Under appropriate conditions known in the art, the transformed plant cells are

grown to form shoots or roots, and develop further into plants. Generally, this procedure involves inoculating the plant tissue with a suspension of bacteria and incubating the tissue for 48 to 72 hours on regeneration medium without antibiotics at 25-28°C.

Methods are explained in various references such as: J. Schell, Science, 237:1176-83 (1987); (U.S. Pat. No. 5,258,300); Herrera-Estrella, Nature, 303, 209 (1983), Biotechnica (published PCT application PCT WO/91/10725), and U.S. patent 4,940,838.

5.7.21 METHOD FOR MAKING GENETICALLY RECOMBINANT PLANTS IN COMMERCIALLY FEASIBLE NUMBERS

In one preferred embodiment, the invention provides for a process for propagating plants by tissue culture in such a way as both to conserve desired plant morphology and to transform the plant with respect to one or more desired genes. The method includes the steps of (a) creating an Agrobacterium vector containing the gene sequence desired to be transferred to the propagated plant, preferably together with a marker gene; (b) taking one or more petiole explants from a mother plant and inoculating them with the Agrobacterium vector; (c) conducting callus formation in the petiole sections in culture, in the dark; and (d) culturing the resulting callus in growth medium having a benzylamino growth regulator such as benzylaminopurine or, most preferably, benzylaminopurineriboside. Additional optional growth regulators including auxins and cytokinins (indole butyric acid, benzylamine, benzyladenine, benzylaminopurine, alpha naphthylacetic acid and others known in the art) may also be present. Preferably, the petiole tissue is taken from Pelargonium x domesticum and the Acrobacterium vector contains an antisense gene for ACC synthase or ACC oxidase to prevent ACC synthase or ACC oxidase expression and, in turn, preventing ethylene formation. Pelargoniums propagated in culture using the present technique are resistant to wilting and petal shatter, and are morphologically conserved due to the use of petiole explants specifically and the particular culture media disclosed. Using a probe for the transposon, the mutated gene can be isolated. Then, using the DNA adjacent to the transposon in the isolated, mutated gene as a probe, the normal wild-type allele of the target gene can be isolated. Such

techniques are taught, for example, in McLaughlin and Walbot, *Genetics*, Vol. 117 pp. 771-776 (1987), as well as numerous other references.

5.7.21.1 Reporter Gene

In addition to the functional gene and the selectable marker gene, the DNA sequences may also contain a reporter gene which facilitates screening of the transformed shoots and plant material for the presence and expression of endogenous DNA sequences. Exemplary reporter genes include β -glucuronidase and luciferase.

5.7.21.2 Transfer Regions of a Suitable Plasmid

As described above, the exogenous DNA sequences are introduced into the area of the explants by incubation with Agrobacterium cells which carry the sequences to be transferred within a transfer DNA (T-DNA) region found on a suitable plasmid, typically the Ti plasmid. Ti plasmids contain two regions essential for the transformation of plant cells. One of these, the T-DNA region, is transferred to the plant nuclei and induces tumor formation. The other, referred to as the virulence (vir) region, is essential for the transfer of the T-DNA but is not itself transferred. By inserting the DNA sequence to be transferred into the T-DNA region, introduction of the DNA sequences to the plant genome can be effected. Usually, the Ti plasmid will be modified to delete or to inactivate the tumor-causing genes so that they are suitable for use as a vector for the transfer of the gene constructs of the present invention. Other plasmids may be utilized in conjunction with Agrobacterium for transferring the DNA sequences of the present invention to the plant cells.

The construction of recombinant Ti plasmids may be accomplished using conventional recombinant DNA techniques, such as those described by Sambrook et al. (1989). Frequently, the plasmids will include additional selective marker genes which

permit manipulation and construction of the plasmids in suitable hosts, typically bacterial hosts other than Agrobacterium, such as E. coli. In addition to the above-described kanamycin resistance marker gene, other exemplary genes are the tetracycline resistance gene and the ampicillin resistance gene, among others.

5.7.21.3 Confirming Transformation

After green transformed shoots are approximately ½" tall, they can then be transplanted to soil within a greenhouse or elsewhere in a conventional manner for tissue culture plantlets. Transformation of the resulting plantlets can be confirmed by assaying activity for the selection marker, or by assaying the plant material for any of the phenotypes which have been introduced by the exogenous DNA. Suitable assay techniques include polymerase chain reaction (PCR), restriction enzyme digestion, Southern blot hybridization and Northern blot hybridization.

5.7.21.4. Commercial Production

The present invention represents a breakthrough in the commercial production and genetically transformed plants. Because the method uses petiole tissue from a grower's mother plant (a stock plant), the starting petiole explants have a commercially desirable morphology to begin with--by definition. However, if the mother plant could be improved by genetic transformation of some type, for example to deactivate a gene which expresses an enzyme in the ethylene synthesis pathway, the progeny of the mother plant may thus be improved in this one way over their parent stock. The petiole tissue from the stock plant, plus the genetic transformation from the Agrobacterium, yield both an improved genetic makeup of the commercially produced plants--although with preserved desired morphology from the mother plant--and at the same time the high yields possible only with the generation of many plantlets in a single generation's growth in tissue culture. In summary, with the present method a single genetically transformed mother

plant can yield literally thousands of offspring plants. No one in the prior art has attempted to combine these two previously disparate technologies to achieve a unique method in which the result is no less than a commercially viable technique for making genetically recombinant plants in commercially feasible numbers.

5.7.22 TRANSFORMATION OF PLANT CELLS USING ALTERNATIVE METHODS

A variety of techniques are available for the introduction of the genetic material into or transformation of the plant cell host. However, the particular manner of introduction of the plant vector into the host is not critical to the practice of the present invention, and any method which provides for efficient transformation may be employed. In addition to transformation using plant transformation vectors derived from the tumor-inducing (Ti) or root-inducing (Ri) plasmids of Agrobacterium, alternative methods could be used to insert the DNA constructs of the present invention into plant cells. Such methods may include, for example, the use of liposomes, transformation using viruses or pollen, chemicals that increase the direct uptake of DNA (Paszkowski et al., EMBO J., 3, 2717 (1984)), microinjection (Crossway et al., Mol. Gen. Genet., 202, 179 (1985)), electroporation (Fromm et al., Proc. Natl. Acad. Sci. US, 82, 824 (1985)), or high-velocity microparticles (Klein et al., Nature, 327, 70 (1987)).

5.7.22.1 Plant Tissue Source or Cultured Plant Cell

The choice of plant tissue source or cultured plant cells for transformation will depend on the nature of the host plant and the - transformation protocol. Useful tissue sources include callus, suspension culture cells, protoplasts, leaf segments, stem segments, tassels, pollen, embryos, hypocotyls, tuber segments, meristematic regions, and the like.

The tissue source is regenerable, in that it will retain the ability to regenerate whole, fertile plants following transformation.

5.7.22.2 Conditions During Transformation

The transformation is carried out under conditions directed to the plant tissue of choice. The plant cells or tissue are exposed to the DNA carrying the present multi-gene expression cassette for an effective period of time. This may range from a less-than-one-second pulse of electricity for electroporation, to a two-to-three day co-cultivation in the presence of plasmid-beazing Agrobacterium cells. Buffers and media used will also vary with the plant tissue source and transformation protocol. Many transformation protocols employ a feeder layer of suspended culture cells (tobacco or Black Mexican Sweet Corn, for example) on the surface of solid media plates, separated by a sterile filter paper disk from the plant cells or tissues being transformed.

Following treatment with DNA, the plant cells or tissue may be cultivated for varying lengths of time prior to selection, or may be immediately exposed to a selective agent such as those described hereinabove.

5.7.22.3 Inhibitory Agent

Protocols involving exposure to Agrobacterium will also include an agent inhibitory to the growth of the Agrobacterium cells. Commonly used compounds are antibiotics such as cefotaxime and carbenicillin. The media used in the selection may be formulated to maintain transformed callus or suspension culture cells in an undifferentiated state, or to allow production of shoots from callus, leaf or stem segments, tuber disks, and the like.

5.7.23 METHOD FOR TRANSFORMATION OF THE TARGET PLANT

The methods used for the actual transformation of the target plant are not critical to this invention. The transformation of the plant is preferably permanent, e.g. by integration of introduced sequences into the plant genome, so that the introduced sequences are passed onto successive plant generations. There are many plant transformation techniques well-known to workers in the art, and new techniques are continually becoming known. Any technique that is suitable for the target plant can be employed with this invention. For example, the sequences can be introduced in a variety of forms, such as a strand of DNA, in a plasmid, or in an artificial chromosome, to name a few. The introduction of the sequences into the target plant cells can be accomplished by a variety of techniques, as well, such as calcium phosphate-DNA co-precipitation, electroporation, microinjection, Agrobacterium infection, liposomes or microprojectile transformation. Those of ordinary skill in the art can refer to the literature for details, and select suitable techniques without undue experimentation.

5.7.23.1 Introduction of Sequences into Target Plant Cells

It is possible to introduce the recombinase gene, in particular, into the transgenic plant in a number of ways. The gene can be introduced along with all of the other basic sequences, as in the first preferred embodiment described above. The repressible promoter/recombinase construct can be also introduced directly via a viral vector into a transgenic plant that contains the other sequence components of the system. Still another method of introducing all the necessary sequences into a single plant is the second preferred embodiment described above, involving a first transgenic plant containing the transiently-active promoter/structural gene sequences and the blocking sequence, and a second transgenic plant containing the recombinase gene linked to a germination-specific plant-active promotor, the two plants being hybridized by conventional to produce hybrid progeny containing all the necessary sequences.

It is also possible to introduce the recombinase itself directly into a transgenic plant as a conjugate with a compound such as biotin, that is transported into the cell. See Horn, et al. (1990).

5.7.4.15.2 Direct or Vectored Transformation

Various methods for direct or vectored transformation of plant cells, e. g., plant protoplast cells, have been described, e.g., in above-cited PCT application WO 95/14099. As noted in that reference, promoters directing expression of selectable markers used for plant transformation (e.g., nptII) should operate effectively in plant hosts. One such promoter is the nos promoter from native Ti plasmids (Heffera-Estrella, et al., *Nature* 303:209-213 (1983). Others include the 35S and 19S promoters of cauliflower mosaic virus (Odell, et al., *Nature* 313:810-812 (1985) and the 2' promoter (Velten, et al., *EMBO J.* 3:2723-2730 (1984).

In one preferred embodiment, the embryo and endosperm. of mature seeds are removed to exposed scutulum tissue cells. The cells may be transformed by DNA bombardment or injection, or by vectored transformation, e.g., by Agrobacteriwn infection after bombarding the scuteller cells with microparticles to make them susceptible to Agrobacteriwn infection (Bidney et al., *Plant Mol. Biol.* 18:301-313, 1992).

One preferred transformation follows the methods detailed generally in Sivamani, E. et al., *Plant Cell Reports* 15:465 (1996); Zhang, S., et al., *Plant Cell Reports* 15:465 (1996); and Li, L., et al., *Plant Cell Reports* 12:250 (1993)..

5.7.24 SUBCULTURING CELLS OR CALLUS GROWING IN NORMALLY INHIBITORY CONCENTRATIONS OF THE SELECTIVE AGENTS

Cells or callus observed to be growing in the presence of normally inhibitory concentrations of the selective agents are presumed to be transformed and may be subcultured several additional times on the same medium to remove non-resistant sections. The cells or calli can then be assayed for the presence of the viral gene cassette,

or may be subjected to known plant regeneration protocols. In protocols involving the direct production of shoots, those shoots appearing on the selective media are presumed to be transformed and may be excised and rooted, either on selective medium suitable for the production of roots, or by simply dipping the excised shoot in a root-inducing compound and directly planting it in vermiculite.

5.7.25 SELECTING FOR MULTI-VIRAL RESISTANCE

In order to produce transgenic plants exhibiting multi-viral resistance, the viral genes must be taken up into the plant cell and stably integrated within the plant genome. Plant cells and tissues selected for their resistance to an inhibitory agent are presumed to have acquired the selectable marker gene encoding this resistance during the transformation treatment.

Since the marker gene is commonly linked to the viral genes, it can be assumed that the viral genes have similarly been acquired. Southern blot hybridization analysis using a probe specific to the viral genes can then be used to confirm that the foreign genes have been taken up and integrated into the genome of the plant cell. This technique may also give some indication of the number of copies of the gene that have been incorporated. Successful transcription of the foreign gene into mRNA can likewise be assayed using Northern blot hybridization analysis of total cellular RNA and/or cellular RNA that has been enriched in a polyadenylated region. mRNA molecules encompassed within the scope of the invention are those which contain viral specific sequences derived from the viral genes present in the transformed vector which are of the same polarity to that of the viral genomic RNA such that they are capable of base pairing with viral specific RNA of the opposite polarity to that of viral genomic RNA under conditions described in Chapter 7 of Sambrook et al. (1989). mRNA molecules also encompassed within the scope of the invention are those which contain viral specific sequences derived from the viral genes present in the transformed vector which are of the opposite polarity

to that of the viral genomic RNA such that they are capable of base pairing with viral genomic RNA under conditions described in Chapter 7 of Sambrook et al. (1989).

The presence of a viral gene can also be detected by immunological assays, such as the double-antibody sandwich assays described by Namba, et al., Gene, 107, 181 (1991) as modified by Clark et al., J. Gen. Virol., 34, 475 (1979). See also, Namba et al., Phytopathology, 82, 940 (1992).

Virus resistance can be assayed via infectivity studies as generally disclosed by Namba et al., ibid., wherein plants are scored as symptomatic when any inoculated leaf shows vein-clearing, mosaic or necrotic symptoms.

It is understood that the invention is operable when either sense or anti-sense viral specific RNA is transcribed from the expression cassettes described above. That is, there is no specific molecular mechanism attributed to the desired phenotype and/or genotype exhibited by the transgenic plants.

Thus, protection against viral challenge can occur by any one or any number of mechanisms.

It is also understood that virus resistance can occur by the expression of any virally encoded gene. Thus, transgenic plants expressing a coat protein gene or a non-coat protein gene can be resistant to challenge with a homologous or heterologous virus..

5.7.26 CELL CULTURE PRODUCTION OF MATURE HETEROLOGOUS PROTEIN

Transgenic cells, typically callus cells, are cultured under conditions that favor plant cell growth, until the cells reach a desired cell density, then under conditions that favor expression of the mature protein under the control of the given promoter. Preferred culture conditions are described herein. Purification of the mature protein secreted into the medium is by standard techniques known by those of skill in the art.

In one embodiment of the invention, in which BPN' is secreted as the proBPN' form of the enzyme, the chaperon "pro" moiety of the enzyme facilitates enzyme folding and is cleaved from the enzyme, leaving the active mature form of BPN'. In another embodiment, the mature enzyme is co-expressed and co-secreted with the "pro" chaperon moiety, with conversion of the enzyme to active form occurring in presence of the free chaperon (Eder et al., Biochem. (1993) L2:18-26; Eder et al, (1993) J. Mol. Biol. 223:293- 304). In yet another embodiment of the invention, the BPN' is secreted in inactive form at a pH that may be in the 6-8 range, with subsequent activation of the inactive form, e.g., after enzyme isolation, by exposure to the "pro" chaperon moiety, e.g., immobilized to a solid support. In both of these embodiments, the culture medium is maintained at a pH of between 5 and 6, preferably about 5.5 during the period of active expression and secretion of BPN', to keep the BPN', which is normally active at alkaline pH, at a pH below optimal activity.

5.7.26.1 Production of Mature Heterologous Protein in Germinatin Seeds

In this embodiment, monocot cells transformed as above are used to regenerate plants, seeds from the plants are harvested and then germinated, and the mature protein is isolated from the germinated seeds.

Plant regeneration from cultured protoplasts or callus tissue is carried by standard methods, e.g., as described in Evans et al., HANDBOOK OF PLANT CELL CULTURE

Vol. 1: (MacMillan Publishing Co. New York, 1983); and Vasil I.R. (ed.), CELL CULTURE AND SOMATIC CELL GENETICS OF PLANTS, Acad. Press, Orlando, Vol. 1, 1984, and Vol. 111, 1986, and as described in the above-cited PCT application.

To achieve maximum production of recombinant protein from malting, the malting procedure may be modified to accommodate de-hulled and de-embryonated seeds, as described in above-cited PCT application WO 95/14099. In the absence of sugars from the endosperm, there is expected to be a 5 to 10 fold increase in RAmy3D promoter activity and thus expression of heterologous protein. Alternatively when embryoless half-seeds are incubated in 10 mM CaCl₂ and 5µM gibberellic acid, there is a 50 fold increase in RAmy1A promoter activity.

5.7.27 REGENERATION OF THE TRANSFORMED PLANT CELLS

After transformation, the transformed plant cells must be regenerated.

The methods used to regenerate transformed cells into whole plants are not critical to this invention, and any method suitable for the target plant can be employed. The literature describes numerous techniques for regenerating specific plant types, (e.g., via somatic embryogenesis, Umbeck, et al., 1987) and more are continually becoming known. Those of ordinary skill in the art can refer to the literature for details and select suitable techniques without undue experimentation.

Plant regeneration from cultured protoplasts is described in Evans et al., Handbook of Plant Cell Cultures, Vol. 1: (MacMillan Publishing Co., New York, 1983); and Vasil I.R. (ed.), Cell Culture and Somatic Cell Genetics of Plants, Acad. Press, Orlando, Vol. I, 1984, and Vol.-III (1986), which are hereby incorporated is by reference.

It is known that practically all plants can be regenerated from cultured cells or tissues, including but not limited to, all major species of sugarcane, sugar beets, cotton, fruit trees, and legumes.

Means for regeneration vary from species to species of plants, but generally a suspension of transformed protoplasts or a petri plate containing explants is first provided. Callus tissue is formed and shoots may be induced from callus and subsequently rooted. Alternatively, embryo formation can be induced in the callus tissue. These embryos germinate as natural embryos to form plants. The culture media will generally contain various amino acids and hormones, such as auxin and cytokinins.. It is also advantageous to add glutamic acid and proline to the medium, especially for such species as corn and alfalfa. Efficient regeneration will depend on the medium, on the genotype, and on the history of the culture. If these three variables are controlled, then regeneration is usually reproducible and repeatable.

5.7.28 BREEDING TECHNIQUES

After the expression cassette is stably incorporated in transgenic plants, it can be transferred to other plants by sexual crossing. Any of a number of standard breeding techniques can be used, depending upon the species to be crossed.

Seed from plants regenerated from tissue culture is grown in the field and self-pollinated to generate true breeding plants. The progeny from these plants become true breeding lines which are evaluated for viral resistance in the field under a range of environmental conditions. The commercial value of viral-resistant plants is greatest if many different hybrid combinations with resistance are available for sale. The farmer typically grows more than one kind of hybrid based on such differences as maturity, disease and insect resistance, color or other agronomic traits. Additionally, hybrids adapted to one part of a country are not adapted to another part because of differences in such traits as maturity, disease and insect tolerance, or public demand for specific varieties in given geographic locations.

Because of this, it is necessary to breed viral resistance into a large number of parental lines so that many hybrid combinations can be produced.

Adding viral resistance to agronomically elite lines is most efficiently accomplished when the genetic control of viral resistance is understood. This requires crossing resistant and sensitive plants and studying the pattern of inheritance in segregating generations to ascertain whether the trait is expressed as dominant or recessive, the number of genes involved, and any possible interaction between genes if more than one are required for expression. With respect to transgenic plants of the type disclosed herein, the transgenes exhibit dominant, single gene Mendelian behavior. This genetic analysis can be part of the initial efforts to convert agronomically elite, yet sensitive lines to resistant lines. A conversion process (backcrossing) is carried out by crossing the original resistant line with a sensitive elite line and crossing the progeny back to the sensitive parent. The progeny from this cross will segregate such that some plants carry the resistance gene(s) whereas some do not. Plants carrying the resistance gene(s) will be crossed again to the sensitive parent resulting in progeny which segregate for resistance and sensitivity once more. This is repeated until the original sensitive parent has been converted to a resistant line, yet possesses all of the other important attributes originally found in the sensitive parent. A separate backcrossing program is implemented for every sensitive elite line that is to be converted to a virus resistant line.

Subsequent to the backcrossing, the new resistant lines and the appropriate combinations of lines which make good commercial hybrids are evaluated for viral resistance, as well as for a battery of important agronomic traits. Resistant lines and hybrids are produced which are true to type of the original sensitive lines and hybrids. This requires evaluation under a range of environmental conditions under which the lines or hybrids will be grown commercially. Parental lines of hybrids that perform satisfactorily are increased and utilized for hybrid production using standard hybrid production practices.

5.7.28.1 USE OF CONVENTIONAL CULTIVATION

Once transgenic plants of this type are produced, the plants themselves can be cultivated in accordance with conventional procedure so that the DNA construct is present in the resulting plants. Alternatively, transgenic seeds are recovered from the transgenic plants. These seeds can then be planted in the soil and cultivated using conventional procedures to produce transgenic plants.

5.7.28.2 PLANT VARIETIES

The present invention can be used to make a variety of transgenic plants. The method is particularly suited for use with plants that are planted as a yearly crop from seed. These include, but are not limited to, fiber crops such as cotton and flax; dicotyledonous seed crops such as soybean, sunflower and peanut; annual ornamental flowers; monocotyledonous grain crops such as maize, wheat and sorghum; leaf crops such as tobacco; vegetable crops such as lettuce, carrot, broccoli, cabbage and cauliflower; and fruit crops such as tomato, zucchini, watermelon, cantaloupe and pumpkin.

The present invention can be utilized in conjunction with a wide variety of plants or their seeds.

Suitable plants include dicots and monocots. More particularly, useful crop plants can include: alfalfa, rice, wheat, barley, rye, cotton, sunflower, peanut, corn, potato, sweet potato, bean, pea, chicory, lettuce, endive, cabbage, brussel sprout, beet, parsnip, turnip, cauliflower, broccoli, turnip, radish, spinach, onion, garlic, eggplant, pepper, celery, carrot, squash, pumpkin, zucchini, cucumber, apple, pear, melon, citrus, strawberry, grape, raspberry, pineapple, soybean, tobacco, tomato, sorghum, papaya, and sugarcane.

Examples of suitable ornamental plants are: *Arabidopsis thaliana*, *Saintpaulia*, *petunia*, *pelargonium*, *poinsettia*, *chrysanthemum*, *carnation*, and *zinnia*.

The plants used in the process of the present invention are derived from monocots, particularly the members of the taxonomic family known as the Gramineae. This family includes all members of the grass family of which the edible varieties are known as cereals. The cereals include a wide variety of species such as wheat (*Triticum* sps.), rice (*Oryza* sps.) barley (*Hordeum* sps.) oats, (*Avena* sps.) rye (*Secale* sps.), corn (*Zea* sps.) and millet (*Pennisetum* sps.). In the present invention, preferred family members are rice and barley.

5.7.29 IDENTIFICATION AND LOCALIZATION AND INTROGRESSION INTO PLANTS OF DESIRED MULTIGENIC TRAITS WITH RFLP TECHNOLOGY

The invention typically involves genetic linkage maps constructed with RFLP technology and the use of RFLP probes to correlate those probes with Quantitative Trait Loci (QTL) and the degree of inheritance of particular multigenic traits (For references see: PTC numbers WO 96/21031; WO 97/17429; WO 98/37223; WO 98/36085; US Pat No. 5,925,808 and 5,385,835).

WHAT IS CLAIMED IS:

1. A method for identifying proteins by differential labeling of peptides, the method comprising the following steps:
 - 5 (a) providing a sample comprising a polypeptide;
 - (b) providing a plurality of labeling reagents which differ in molecular mass that can generate differential labeled peptides that do not differ in chromatographic retention properties and do not differ in ionization and detection properties in mass spectrographic analysis, wherein the differences in molecular mass are distinguishable by mass spectrographic analysis;
 - 10 (c) fragmenting the polypeptide into peptide fragments by enzymatic digestion or by non-enzymatic fragmentation;
 - (d) contacting the labeling reagents of step (b) with the peptide fragments of step (c), thereby labeling the peptides with the differential labeling reagents;
 - (e) separating the peptides by chromatography to generate an eluate;
 - 15 (f) feeding the eluate of step (e) into a mass spectrometer and quantifying the amount of each peptide and generating the sequence of each peptide by use of the mass spectrometer;
 - (g) inputting the sequence to a computer program product which compares the inputted sequence to a database of polypeptide sequences to identify the polypeptide from which the sequenced peptide originated.
- 20
25 2. The method of claim 1, wherein the sample of step (a) comprises a cell or a cell extract.
- 30 3. The method of claim 1, further comprising providing two or more samples comprising a polypeptide.
4. The method of claim 3, wherein one sample is derived from a wild type cell and one sample is derived from an abnormal or a modified cell.
- 35 5. The method of claim 4, wherein the abnormal cell is a cancer cell.
6. The method of claim 1, further comprising purifying or fractionating the polypeptide before the fragmenting of step (c).

7. The method of claim 1, further comprising purifying or fractionating the polypeptide before the labeling of step (d).

5 8. The method of claim 1, further comprising purifying or fractionating the labeled peptide before the chromatography of step (e).

10 9. The method of claim 6, claim 8 or claim 8, wherein the purifying or fractionating comprises a method selected from the group consisting of size exclusion chromatography, size exclusion chromatography, HPLC, reverse phase HPLC and affinity purification.

15 10. The method of claim 1, further comprising contacting the polypeptide with a labeling reagent of step (b) before the fragmenting of step (c).

15 11. The method of claim 1, wherein the labeling reagent of step (b) comprises the general formulae selected from the group consisting of:

i. $Z^A\text{OH}$ and $Z^B\text{OH}$, to esterify peptide C-terminals and/or Glu and Asp side chains;

20 ii. $Z^A\text{NH}_2$ and $Z^B\text{NH}_2$, to form amide bond with peptide C-terminals and/or Glu and Asp side chains; and

iii. $Z^A\text{CO}_2\text{H}$ and $Z^B\text{CO}_2\text{H}$, to form amide bond with peptide N-terminals and/or Lys and Arg side chains;

25 wherein Z^A and Z^B independently of one another comprise the general formula $R-Z^1-A^1-Z^2-A^2-Z^3-A^3-Z^4-A^4-$,

25 Z^1, Z^2, Z^3 , and Z^4 independently of one another, are selected from the group consisting of nothing, O, OC(O), OC(S), OC(O)O, OC(O)NR, OC(S)NR, OSiRR¹, S, SC(O), SC(S), SS, S(O), S(O)₂, NR, NRR¹⁺, C(O), C(O)O, C(S), C(S)O, C(O)S, C(O)NR, C(S)NR, SiRR¹, (Si(RR¹)O)n, SnRR¹, Sn(RR¹)O, BR(OR¹), BRR¹, B(OR)(OR¹), OBR(OR¹), OBRR¹, and OB(OR)(OR¹), and R and R¹ is an alkyl group,

30 A^1, A^2, A^3 , and A^4 independently of one another, are selected from the group consisting of nothing or (CRR¹)n, wherein R, R¹, independently from other R and R¹ in Z^1 to Z^4 and independently from other R and R¹ in A^1 to A^4 , are selected from the group consisting of a hydrogen atom, a halogen atom and an alkyl group;

35 n in Z^1 to Z^4 , independent of n in A^1 to A^4 , is an integer having a value selected from the group consisting of 0 to about 51; 0 to about 41; 0 to about 31; 0 to about 21, 0 to about 11 and 0 to about 6.

40 12. The method of claim 11, wherein the alkyl group is selected from the group consisting of an alkenyl, an alkynyl and an aryl group.

13. The method of claim 11, wherein one or more C-C bonds from (CRR¹)n are replaced with a double or a triple bond,

14. The method of claim 13, wherein an R or an R¹ group is deleted.

15. The method of claim 13, wherein $(CRR^1)_n$ is selected from the group consisting of an *o*-arylene, an *m*-arylene and a *p*-arylene, wherein each group has none or up to 6 substituents.

5

16. The method of claim 13, wherein $(CRR^1)_n$ is selected from the group consisting of a carbocyclic, a bicyclic and a tricyclic fragment, wherein the fragment has up to 8 atoms in the cycle with or without a heteroatom selected from the group consisting of an O atom, a N atom and an S atom.

10

17. The method of claim 1, wherein two or more labeling reagents have the same structure but a different isotope composition.

15

18. The method of claim 11, wherein Z^A has the same structure as Z^B , but Z^A has a different isotope composition than Z^B .

19. The method of claim 17, wherein the isotope is boron-10 and boron-11.

20

20. The method of claim 17, wherein the isotope is carbon-12 and carbon-13.

21. The method of claim 17, wherein the isotope is nitrogen-14 and nitrogen-15.

25

22. The method of claim 17, wherein the isotope is sulfur-32 and sulfur-34.

30

23. The method of claim 17, wherein, where the isotope with the lower mass is x and the isotope with the higher mass is y, and x and y are integers, x is greater than y.

35

24. The method of claim 17, wherein x and y are between 1 and about 11, between 1 and about 21, between 1 and about 31, between 1 and about 41, or between 1 and about 51.

25. The method of claim 1, wherein the labeling reagent of step (b) comprises the general formulae selected from the group consisting of:

40

i. $CD_3(CD_2)_nOH / CH_3(CH_2)_nOH$, to esterify peptide C-terminals, where $n = 0, 1, 2$ or y ;

ii. $CD_3(CD_2)_nNH_2 / CH_3(CH_2)_nNH_2$, to form amide bond with peptide C-terminals, where $n = 0, 1, 2$ or y ; and

iii. $D(CD_2)_nCO_2H / H(CH_2)_nCO_2H$, to form amide bond with peptide N-terminals, where $n = 0, 1, 2$ or y ;

wherein D is a deuteron atom, and y is an integer selected from the group consisting of about 51; about 41; about 31; about 21, about 11; about 6 and between about 5 and 51.

5 26. The method of claim 1, wherein the labeling reagent of step (b) comprises the general formulae selected from the group consisting of:

- i. $Z^A\text{OH}$ and $Z^B\text{OH}$ to esterify peptide C-terminals;
- ii. $Z^A\text{NH}_2$ / $Z^B\text{NH}_2$ to form an amide bond with peptide C-terminals; and
- iii. $Z^A\text{CO}_2\text{H}$ / $Z^B\text{CO}_2\text{H}$ to form an amide bond with peptide N-terminals;

10 wherein Z^A and Z^B have the general formula $R\text{-}Z^1\text{-}A^1\text{-}Z^2\text{-}A^2\text{-}Z^3\text{-}A^3\text{-}Z^4\text{-}A^4\text{-}$

15 Z^1, Z^2, Z^3 , and Z^4 , independently of one another, are selected from the group consisting of nothing, O, OC(O), OC(S), OC(O)O, OC(O)NR, OC(S)NR, OSiRR¹, S, SC(O), SC(S), SS, S(O), S(O₂), NR, NRR¹⁺, C(O), C(O)O, C(S), C(S)O, C(O)S, C(O)NR, C(S)NR, SiRR¹, (Si(RR¹)O)n, SnRR¹, Sn(RR¹)O, BR(OR¹), BRR¹,

B(OR)(OR¹), OBR(OR¹), OBRR¹, and OB(OR)(OR¹);

20 A^1, A^2, A^3 , and A^4 , independently of one another, are selected from the group consisting of nothing and the general formulae (CRR¹)n, and,

R and R^1 is an alkyl group.

25 27. The method of claim 26, wherein a single C-C bond in a (CRR¹)n group is replaced with a double or a triple bond.

30 28. The method of claim 27, wherein R and R¹ are absent.

35 29. The method of claim 27, wherein (CRR¹)n comprises a moiety selected from the group consisting of an *o*-arylene, an *m*-arylene and a *p*-arylene, wherein the group has none or up to 6 substituents.

40 30. The method of claim 27, wherein the group comprises a carbocyclic, a bicyclic, or a tricyclic fragments with up to 8 atoms in the cycle, with or without a heteroatom selected from the group consisting of an O atom, an N atom and an S atom.

35 31. The method of claim 26, wherein R, R¹, independently from other R and R¹ in $Z^1 - Z^4$ and independently from other R and R¹ in $A^1 - A^4$, are selected from the group consisting of a hydrogen atom, a halogen and an alkyl group.

40 32. The method of claim 31, wherein the alkyl group is selected from the group consisting of an alkenyl, an alkynyl and an aryl group.

35 33. The method of claim 26, wherein n in $Z^1 - Z^4$ is independent of n in $A^1 - A^4$ and is an integer selected from the group consisting of about 51; about 41; about 31; about 21, about 11 and about 6.

34. The method of claim 26, wherein Z^A has the same structure a Z^B but Z^A further comprises x number of - CH_2- fragment(s) in one or more $A^1 - A^4$ fragments, wherein x is an integer.

5 35. The method of claim 26, wherein Z^A has the same structure a Z^B but Z^A further comprises x number of - CF_2- fragment(s) in one or more $A^1 - A^4$ fragments, wherein x is an integer.

10 36. The method of claim 26, wherein Z^A comprises x number of protons and Z^B comprises y number of halogens in the place of protons, wherein x and y are integers.

15 37. The method of claim 26, wherein Z^A contains x number of protons and Z^B contains y number of halogens, and there are $x - y$ number of protons remaining in one or more $A^1 - A^4$ fragments, wherein x and y are integers

20 38. The method of claim 26, wherein Z^A further comprises x number of - $\text{O}-$ fragment(s) in one or more $A^1 - A^4$ fragments, wherein x is an integer.

25 39. The method of claim 26, wherein Z^A further comprises x number of - $\text{S}-$ fragment(s) in one or more $A^1 - A^4$ fragments, wherein x is an integer.

40. The method of claim 26, wherein Z^A further comprises x number of - $\text{O}-$ fragment(s) and Z^B further comprises y number of - $\text{S}-$ fragment(s) in the place of - $\text{O}-$ fragment(s), wherein x and y are integers.

30 41. The method of claim 26, wherein Z^A further comprises $x - y$ number of - $\text{O}-$ fragment(s) in one or more $A^1 - A^4$ fragments, wherein x and y are integers.

35 42. The method of claim 37, claim 40 or claim 41, wherein x and y are integers selected from the group consisting of between 1 about 51; between 1 about 41; between 1 about 31; between 1 about 21, between 1 about 11 and between 1 about 6, wherein x is greater than y .

43. The method of claim 1, wherein the labeling reagent of step (b) comprises the general formulae selected from the group consisting of:

- i. $\text{CH}_3(\text{CH}_2)_n\text{OH}/\text{CH}_3(\text{CH}_2)_{n+m}\text{OH}$, to esterify peptide C-terminals, where $n = 0, 1, 2, \dots, y$; $m = 1, 2, \dots, y$;
- ii. $\text{CH}_3(\text{CH}_2)_n\text{NH}_2 / \text{CH}_3(\text{CH}_2)_{n+m}\text{NH}_2$, to form amide bond with peptide C-terminals, where $n = 0, 1, 2, \dots, y$; $m = 1, 2, \dots, y$; and,
- iii. $\text{H}(\text{CH}_2)_n\text{CO}_2\text{H} / \text{H}(\text{CH}_2)_{n+m}\text{CO}_2\text{H}$, to form amide bond with peptide N-terminals, where $n = 0, 1, 2, \dots, y$; $m = 1, 2, \dots, y$;
wherein n, m and y are integers.

44. The method of claim 43, wherein n, m and y are integers selected from the group consisting of about 51; about 41; about 31; about 21, about 11; about 6 and between about 5 and 51.

5 45. The method of claim 1, wherein the separating of step (e) comprises a liquid chromatography system.

10 46. The method of claim 1, wherein the liquid chromatography system comprises a multidimensional liquid chromatography.

15 47. The method of claim 1, wherein the mass spectrometer comprises a tandem mass spectrometry device.

48. The method of claim 1, further comprising quantifying the amount 15 of each polypeptide.

49. The method of claim 1, further comprising quantifying the amount of each peptide.

20 50. A method for defining the expressed proteins associated with a given cellular state, the method comprising the following steps:

(a) providing a sample comprising a cell in the desired cellular state;

25 (b) providing a plurality of labeling reagents which differ in molecular mass but do not differ in chromatographic retention properties and do not differ in ionization and detection properties in mass spectrographic analysis, wherein the differences in molecular mass are distinguishable by mass spectrographic analysis;

(c) fragmenting polypeptides derived from the cell into peptide fragments by enzymatic digestion or by non-enzymatic fragmentation;

30 (d) contacting the labeling reagents of step (b) with the peptide fragments of step (c), thereby labeling the peptides with the differential labeling reagents;

(e) separating the peptides by chromatography to generate an eluate;

(f) feeding the eluate of step (e) into a mass spectrometer and quantifying the amount of each peptide and generating the sequence of each peptide by use of the mass spectrometer;

35 (g) inputting the sequence to a computer program product which compares the inputted sequence to a database of polypeptide sequences to identify the polypeptide from which the sequenced peptide originated, thereby defining the expressed proteins associated with the cellular state.

40 51. A method for quantifying changes in protein expression between at least two cellular states, the method comprising the following steps:

- (a) providing at least two samples comprising cells in a desired cellular state;
- (b) providing a plurality of labeling reagents which differ in molecular mass but do not differ in chromatographic retention properties and do not differ in ionization and detection properties in mass spectrographic analysis, wherein the differences in molecular mass are distinguishable by mass spectrographic analysis;
- (c) fragmenting polypeptides derived from the cells into peptide fragments by enzymatic digestion or by non-enzymatic fragmentation;
- (d) contacting the labeling reagents of step (b) with the peptide fragments of step (c), thereby labeling the peptides with the differential labeling reagents, wherein the labels used in one sample are different from the labels used in other samples;
- (e) separating the peptides by chromatography to generate an eluate;
- (f) feeding the eluate of step (e) into a mass spectrometer and quantifying the amount of each peptide and generating the sequence of each peptide by use of the mass spectrometer;
- (g) inputting the sequence to a computer program product which identifies from which sample each peptide was derived, compares the inputted sequence to a database of polypeptide sequences to identify the polypeptide from which the sequenced peptide originated, and compares the amount of each polypeptide in each sample, thereby quantifying changes in protein expression between at least two cellular states.
52. A method for identifying proteins by differential labeling of peptides, the method comprising the following steps:
- (a) providing a sample comprising a polypeptide;
- (b) providing a plurality of labeling reagents which differ in molecular mass but do not differ in chromatographic retention properties and do not differ in ionization and detection properties in mass spectrographic analysis, wherein the differences in molecular mass are distinguishable by mass spectrographic analysis;
- (c) fragmenting the polypeptide into peptide fragments by enzymatic digestion or by non-enzymatic fragmentation;
- (d) contacting the labeling reagents of step (b) with the peptide fragments of step (c), thereby labeling the peptides with the differential labeling reagents;
- (e) separating the peptides by multidimensional liquid chromatography to generate an eluate;

- (f) feeding the eluate of step (e) into a tandem mass spectrometer and quantifying the amount of each peptide and generating the sequence of each peptide by use of the mass spectrometer;
- (g) inputting the sequence to a computer program product which
- 5 compares the inputted sequence to a database of polypeptide sequences to identify the polypeptide from which the sequenced peptide originated.
- 10 53. A chimeric labeling reagent comprising
(a) a first domain comprising a biotin; and
(b) a second domain comprising a reactive group capable of covalently binding to an amino acid,
wherein the chimeric labeling reagent comprises at least one isotope.
- 15 54. The chimeric labeling reagent of claim 53, wherein the isotope is in the first domain.
55. The chimeric labeling reagent of claim 54, wherein the isotope is in the biotin.
- 20 56. The chimeric labeling reagent of claim 53, wherein the isotope is in the second domain.
- 25 57. The chimeric labeling reagent of claim 53, wherein the isotope is selected from the group consisting of a deuterium isotope, a boron-10 or boron-11 isotope, a carbon-12 or a carbon-13 isotope, a nitrogen-14 or a nitrogen-15 isotope and a sulfur-32 or a sulfur-34 isotope.
- 30 58. The chimeric labeling reagent of claim 53 comprising two or more isotopes.
59. The chimeric labeling reagent of claim 53, wherein the reactive group capable of covalently binding to an amino acid is selected from the group consisting of a succimide group, an isothiocyanate group and an isocyanate group.
- 35 60. The chimeric labeling reagent of claim 53, wherein the reactive group capable of covalently binding to an amino acid binds to a lysine or a cysteine.
- 40 61. The chimeric labeling reagent of claim 53, further comprising a linker moiety linking the biotin group and the reactive group.
62. The chimeric labeling reagent of claim 53; wherein the linker moiety comprises at least one isotope.

63. The chimeric labeling reagent of claim 53, wherein the linker is a cleavable moiety.

5 64. The chimeric labeling reagent of claim 53, wherein the linker can be cleaved by enzymatic digest.

65. The chimeric labeling reagent of claim 53, wherein the linker can be cleaved by reduction.

10 66. A method of comparing relative protein concentrations in a sample comprising

(a) providing a plurality of differential small molecule tags, wherein the small molecule tags are structurally identical but differ in their isotope composition, and the small molecules comprise reactive groups that covalently bind to cysteine or lysine residues or both;

(b) providing at least two samples comprising polypeptides;

(c) attaching covalently the differential small molecule tags to amino acids of the polypeptides;

20 (d) determining the protein concentrations of each sample in a tandem mass spectrometer; and,

(d) comparing relative protein concentrations of each sample.

25 67. The method of claim 66, wherein the sample comprises a complete or a fractionated cellular sample.

30 68. The method of claim 66, wherein differential small molecule tags comprise a chimeric labeling reagent comprising (a) a first domain comprising a biotin; and, (b) a second domain comprising a reactive group capable of covalently binding to an amino acid, wherein the chimeric labeling reagent comprises at least one isotope.

35 69. The method of claim 68, wherein the isotope is selected from the group consisting of a deuterium isotope, a boron-10 or boron-11 isotope, a carbon-12 or a carbon-13 isotope, a nitrogen-14 or a nitrogen-15 isotope and a sulfur-32 or a sulfur-34 isotope.

35 70. The method of claim 68, wherein the chimeric labeling reagent comprises two or more isotopes.

40 71. The method of claim 68, wherein the reactive group capable of covalently binding to an amino acid is selected from the group consisting of a succimide group, an isothiocyanate group and an isocyanate group.

72. A method of comparing relative protein concentrations in a sample comprising
- (a) providing a plurality of differential small molecule tags, wherein the differential small molecule tags comprise a chimeric labeling reagent comprising (i) a first domain comprising a biotin; and, (ii) a second domain comprising a reactive group capable of covalently binding to an amino acid, wherein the chimeric labeling reagent comprises at least one isotope;
- (b) providing at least two samples comprising polypeptides;
- (c) attaching covalently the differential small molecule tags to amino acids of the polypeptides;
- (d) isolating the tagged polypeptides on a biotin-binding column by binding tagged polypeptides to the column, washing non-bound materials off the column, and eluting tagged polypeptides off the column;
- (e) determining the protein concentrations of each sample in a tandem mass spectrometer; and,
- (f) comparing relative protein concentrations of each sample.

5 **NOVEL METHODS & REAGENTS FOR CELLULAR & METABOLIC
ENGINEERING, HOLISTIC MONITORING, AND REAL-TIME FLUX
ANALYSIS**

10 Second set of claims

- 15 73. A method of producing an improved organism having a desirable trait comprising: a) obtaining an initial population of organisms, b) generating a set of mutagenized organisms, such that when all the genetic mutations in the set of mutagenized organisms are taken as a whole, there is represented a set of substantial genetic mutations, and c) detecting the presence of said improved organism.
- 20 74. The method of claim 73, wherein the set of substantial genetic mutations in step b) is comprised of a knocking out of at least 15 different genes.
- 25 75. The method of claim 73, wherein the set of substantial genetic mutations in step b) is comprised of a knocking out of at least 50 different genes.
76. The method of claim 73, wherein the set of substantial genetic mutations in step b) is comprised of a knocking out of at least 100 different genes.
- 25 77. The method of claim 73, wherein the set of substantial genetic mutations in step b) is comprised of an introduction of at least 15 different genes.
- 30 78. The method of claim 73, wherein the set of substantial genetic mutations in step b) is comprised of an introduction of at least 50 different genes.
79. The method of claim 73, wherein the set of substantial genetic mutations in step b) is comprised of an introduction of at least 100 different genes.

80. The method of claim 73, wherein the set of substantial genetic mutations in step
b) is comprised of an alteration in the expression of at least 15 different genes.
- 5 81. The method of claim 73, wherein the set of substantial genetic mutations in step
b) is comprised of an alteration in the expression of at least 50 different genes.
- 10 82. The method of claim 73, wherein the set of substantial genetic mutations in step
b) is comprised of an alteration in the expression of at least 100 different genes.
- 15 83. A method of producing an improved organism having a desirable trait
comprising: a) obtaining an initial population of organisms, b) generating a set of
mutagenized organisms each having at least one genetic mutation, such that when
all the genetic mutations in the set of mutagenized organisms are taken as a
whole, there is represented a set of substantial genetic mutations c) detecting the
manifestation of at least two genetic mutations, d) introducing at least two
detected genetic mutations into one organism, and e) optionally repeating any of
steps a), b), c), and d).
- 20 84. The method of claim 83, wherein step d) is comprised of a knocking out of at
least 15 different genes in one organism.
- 25 85. The method of claim 83, wherein step d) is comprised of a knocking out of at
least 50 different genes in one organism.
86. The method of claim 83, wherein step d) is comprised of a knocking out of at
least 100 different genes in one organism.
- 30 87. The method of claim 83, wherein step d) is comprised of an introduction of at
least 15 different genes into one organism.

88. The method of claim 83, wherein step d) is comprised of an introduction of at least 50 different genes into one organism.
- 5 89. The method of claim 83, wherein step d) is comprised of an introduction of at least 100 different genes into one organism.
90. The method of claim 83, wherein step d) is comprised of an alteration in the expression of at least 15 different genes in one organism.
- 10 91. The method of claim 83, wherein step d) is comprised of an alteration in the expression of at least 50 different genes in one organism.
92. The method of claim 83, wherein step d) is comprised of an alteration in the expression of at least 100 different genes in one organism.
- 15 93. A method for identifying a gene that alters a trait of an organism, comprising: a) obtaining an initial population of organisms, b) generating a set of mutagenized organisms, such that when all the genetic mutations in the set of mutagenized organisms are taken as a whole, there is represented a set of substantial genetic mutations, and c) detecting the presence of an organism having said altered trait, and d) determining the nucleotide sequence of a gene that has been mutagenized in the organism having the altered trait.
- 20 94. A method for producing an organism with an improved trait, comprising: a) functionally knocking out an endogenous gene in a substantially clonal population of organisms; b) transferring a library of altered genes into the substantially clonal population of organisms, wherein each altered gene differs from the endogenous gene at only one codon; c) detecting a mutagenized organism having an improved trait; and d) determining the nucleotide sequence of a gene that has been transferred into the detected organism.
- 25
- 30

95. A method of introducing differentially activatable stacked traits into a transgenic cell or organism, which method is comprised of the following steps:
- a) obtaining an initial cell or organism;
 - b) introducing into the working cell or organism a plurality of traits (stacked traits), including selectively and differentially activatable traits, whereby serviceable traits for this purpose include traits conferred by genes and traits conferred by gene pathways;
 - c) analyzing the information obtained from steps a) and b), and
 - d) optionally repeating any number or all of the steps of a), b), c), and d);
- 10 96. The method of Claim 95, wherein step a) also includes holistic monitoring of the strain or organism whereby holistic monitoring can include the detection and/or measurement of all detectable functions and physical parameters (such as but not limited to morphology, behavior, growth, responsiveness to stimuli [e.g., antibiotics, different environment, etc.], and profiles of all detectable molecules, including molecules that are chemically at least in part a nucleic acids, proteins, carbohydrates, proteoglycans, glycoproteins, or lipids)
- 20 97. The method of Claim 95, wherein step d) also includes holistic monitoring of the strain or organism whereby holistic monitoring can include the detection and/or measurement of all detectable functions and physical parameters (such as but not limited to morphology, behavior, growth, responsiveness to stimuli [e.g., antibiotics, different environment, etc.], and profiles of all detectable molecules, including molecules that are chemically at least in part a nucleic acids, proteins, carbohydrates, proteoglycans, glycoproteins, or lipids)
- 25 98. The method of Claim 95, wherein step a) and d) include holistic monitoring of the strain or organism whereby holistic monitoring can include the detection and/or measurement of all detectable functions and physical parameters (such as but not limited

to morphology, behavior, growth, responsiveness to stimuli [e.g., antibiotics, different environment, etc.], and profiles of all detectable molecules, including molecules that are chemically at least in part a nucleic acids, proteins, carbohydrates, proteoglycans, glycoproteins, or lipids)

5

99. The method of Claim 95, wherein step b) includes the introduction of at least 15 stacked traits

10 100. The method of Claim 95, wherein step b) includes the introduction of at least 50 stacked traits

15 101. The method of Claim 95, wherein step b) includes the introduction of at least 100 stacked traits

102. The method of Claim 96, wherein step a) includes screening cellular characteristics by utilizing one or any combination of the following methods:

- 20 a) genomics;
b) transcriptome characterization or RNA profiling;
c) proteomics;
d) metabolomics or the analysis of metabolites;
e) lipidomics or lipid profiling.

25 103. A method of Claim 102, wherein proteomics specifically includes the use of amino acid reactive tags

30 104. A method of Claim 97, wherein step d) includes screening cellular characteristics by utilizing one or any combination of the following methods:

- f) genomics;
g) transcriptome characterization or RNA profiling;
h) proteomics;
i) metabolomics or the analysis of metabolites;
j) lipidomics or lipid profiling.

35 105. A method of Claim 104, wherein proteomics specifically includes the use of amino acid reactive tags

106. A method of Claim 98, wherein steps a) and d) include screening cellular characteristics by utilizing one or any combination of the following methods:
- k) genomics;
l) transcriptome characterization or RNA profiling;
5 m) proteomics;
n) metabolomics or the analysis of metabolites;
o) lipidomics or lipid profiling.
p)
107. A method of Claim 106, wherein proteomics specifically includes the use of amino acid reactive tags
108. A method of Claim 73, wherein step c) includes screening cellular characteristics by utilizing one or any combination of the following methods:
- q) genomics;
15 r) transcriptome characterization or RNA profiling;
s) proteomics;
t) metabolomics or the analysis of metabolites;
u) lipidomics or lipid profiling.
- 20 109. A method of Claim 108, wherein proteomics specifically includes the use of amino acid reactive tags
110. A method of Claim 93, wherein step c) includes screening cellular characteristics by utilizing one or any combination of the following methods:
- v) genomics;
w) transcriptome characterization or RNA profiling;
25 x) proteomics;
y) metabolomics or the analysis of metabolites;
z) lipidomics or lipid profiling.
- 30 111. A method of Claim 110, wherein proteomics specifically includes the use of amino acid reactive tags
112. A method of Claim 94, wherein step c) includes screening cellular characteristics by utilizing one or any combination of the following methods:
- aa) genomics;
bb) transcriptome characterization or RNA profiling;
cc) proteomics;
35 dd) metabolomics or the analysis of metabolites;
ee) lipidomics or lipid profiling.
- 40 113. A method of Claim 112, wherein proteomics specifically includes the use of amino acid reactive tags
- 45 114. A method for whole cell engineering of new or modified phenotypes by using

real-time metabolic flux analysis, the method comprising the following steps:

- (a) making a modified cell by modifying the genetic composition of a cell;
 - (b) culturing the modified cell to generate a plurality of modified cells;
 - (c) measuring at least one metabolic parameter of the cell by monitoring the cell culture of step (b) in real time; and,
 - (d) analyzing the data of step (c) to determine if the measured parameter differs from a comparable measurement in an unmodified cell under similar conditions, thereby identifying an engineered phenotype in the cell using real-time metabolic flux analysis.

10

115. The method of claim 114, wherein the genetic composition of the cell is modified by a method comprising addition of a nucleic acid to the cell.

15

116. The method of claim 115, wherein the nucleic acid comprises a nucleic acid heterologous to the cell.

117. The method of claim 115, wherein the nucleic acid comprises a nucleic acid homologous to the cell.

118. The method of claim 117, wherein the homologous nucleic acid comprises
a modified homologous nucleic acid.

25

119. The method of claim 118, wherein the homologous nucleic acid comprises
a modified homologous gene.

30

120. The method of claim 114, wherein the genetic composition of the cell is modified by a method comprising deletion of a sequence or modification of a sequence in the cell.

121. The method of claim 114, wherein the genetic composition of the cell is modified by a method comprising modifying or knocking out the expression of a gene.
- 5 122. The method of claim 114, further comprising selecting a cell comprising a newly engineered phenotype.
- 10 123. The method of claim 122, further comprising culturing the selected cell, thereby generating a new cell strain comprising a newly engineered phenotype.
- 15 124. The method of claim 122, wherein the newly engineered phenotype is selected from the group consisting of an increased or decreased expression or amount of a polypeptide, an increased or decreased amount of an mRNA transcript, an increased or decreased expression of a gene, an increased or decreased resistance or sensitivity to a toxin, an increased or decreased resistance use or production of a metabolite, an increased or decreased uptake of a compound by the cell, an increased or decreased rate of metabolism, and an increased or decreased growth rate.
- 20 125. The method of claim 114, further comprising isolating a cell comprising a newly engineered phenotype.
126. The method of claim 114, wherein the newly engineered phenotype is a stable phenotype.
- 25 127. The method of claim 126, wherein modifying the genetic composition of a cell comprises insertion of a construct into the cell, wherein construct comprises a nucleic acid operably linked to a constitutively active promoter.
- 30 128. The method of claim 114, wherein the newly engineered phenotype is an inducible phenotype.

129. The method of claim 128, wherein modifying the genetic composition of a cell comprises insertion of a construct into the cell, wherein construct comprises a nucleic acid operably linked to an inducible promoter.

5

130. The method of claim 115, wherein nucleic acid added to the cell in step (a) is stably inserted into the genome of the cell.

10

131. The method of claim 115, wherein nucleic acid added to the cell in step (a) propagates as an episome in the cell.

132. The method of claim 115, wherein nucleic acid added to the cell in step (a) encodes a polypeptide.

15

133. The method of claim 132, wherein the polypeptide comprises a modified homologous polypeptide.

134. The method of claim 132, wherein the polypeptide comprises a heterologous polypeptide.

20

135. The method of claim 115, wherein the nucleic acid added to the cell in step (a) encodes a transcript comprising a sequence that is antisense to a homologous transcript.

25

136. The method of claim 114, wherein modifying the genetic composition of the cell in step (a) comprises increasing or decreasing the expression of an mRNA transcript.

30

137. The method of claim 114, wherein modifying the genetic composition of the cell in step (a) comprises increasing or decreasing the expression of a polypeptide.

138. The method of claim 114, wherein modifying the homologous gene in step
(a) comprises knocking out expression of the homologous gene.

5 139. The method of claim 114, wherein modifying the homologous gene in step
(a) comprises increasing the expression of the homologous gene.

10 140. The method of claim 114, wherein the heterologous gene in step (a)
comprises a sequence-modified homologous gene, wherein the sequence
modification is made by a method comprising the following steps:
(a) providing a template polynucleotide, wherein the template polynucleotide
comprises a homologous gene of the cell;
(b) providing a plurality of oligonucleotides, wherein each oligonucleotide
comprises a sequence homologous to the template polynucleotide, thereby targeting
15 a specific sequence of the template polynucleotide, and a sequence that is a variant
of the homologous gene;
(c) generating progeny polynucleotides comprising non-stochastic sequence
variations by replicating the template polynucleotide of step (a) with the
oligonucleotides of step (b), thereby generating polynucleotides comprising
20 homologous gene sequence variations.

25 141. The method of claim 114, wherein the heterologous gene in step (a)
comprises a sequence-modified homologous gene, wherein the sequence
modification is made by a method comprising the following steps:
(a) providing a template polynucleotide, wherein the template polynucleotide
comprises sequence encoding a homologous gene;
(b) providing a plurality of building block polynucleotides, wherein the building
block polynucleotides are designed to cross-over reassemble with the template
polynucleotide at a predetermined sequence, and a building block polynucleotide
30 comprises a sequence that is a variant of the homologous gene and a sequence
homologous to the template polynucleotide flanking the variant sequence;

(c) combining a building block polynucleotide with a template polynucleotide such that the building block polynucleotide cross-over reassembles with the template polynucleotide to generate polynucleotides comprising homologous gene sequence variations.

5

142. The method of claim 114, wherein the cell is a prokaryotic cell.

143. The method of claim 142, wherein the prokaryotic cell is a bacterial cell.

10

144. The method of claim 114, wherein the cell is selected from the group consisting of a fungal cell, a yeast cell, a plant cell and an insect cell.

145. The method of claim 114, wherein the cell is a eukaryotic cell.

15

146. The method of claim 145, wherein the cell is a mammalian cell.

147. The method of claim 146, wherein the mammalian cell is a human cell.

20

148. The method of claim 114, wherein the measured metabolic parameter comprises rate of cell growth.

149. The method of claim 148, wherein the rate of cell growth is measured by a change in optical density of the culture.

25

150. The method of claim 114, wherein the measured metabolic parameter comprises a change in the expression of a polypeptide.

30

151. The method of claim 150, wherein the change in the expression of the polypeptide is measured by a method selected from the group consisting of a one-dimensional gel electrophoresis, a two-dimensional gel electrophoresis, a tandem mass spectography, an RIA, an ELISA, an immunoprecipitation and a Western blot.

152. The method of claim 114, wherein the measured metabolic parameter comprises a change in expression of at least one transcript, or, the expression of a transcript of a newly introduced gene.

5

153. The method of claim 152, wherein the change in expression of the transcript is measured by a method selected from the group consisting of a hybridization, a quantitative amplification and a Northern blot.

10

154. The method of claim 153, wherein transcript expression is measured by hybridization of a sample comprising transcripts of a cell or nucleic acid representative of or complementary to transcripts of a cell by hybridization to immobilized nucleic acids on an array.

15

155. The method of claim 114, wherein the measured metabolic parameter comprises an increase or a decrease in a secondary metabolite.

156. The method of claim 155, wherein secondary metabolite is selected from the group consisting of a glycerol and a methanol.

20

157. The method of claim 114, wherein the measured metabolic parameter comprises an increase or a decrease in an organic acid.

25

158. The method of claim 157, wherein the organic acid is selected from the group consisting of an acetate, a butyrate, a succinate and an oxaloacetate.

159. The method of claim 114, wherein the measured metabolic parameter comprises an increase or a decrease in intracellular pH.

30

160. The method of claim 159, wherein the increase or a decrease in intracellular pH is measured by intracellular application of a dye, and the change in

fluorescence of the dye is measured over time.

161. The method of claim 114, wherein the measured metabolic parameter comprises an increase or a decrease in synthesis of DNA over time.

5

162. The method of claim 161, wherein the increase or a decrease in synthesis of DNA over time is measured by intracellular application of a dye, and the change in fluorescence of the dye is measured over time.

10

163. The method of claim 114, wherein the measured metabolic parameter comprises an increase or a decrease in uptake of a composition.

164. The method of claim 163, wherein the composition is a metabolite.

15

165. The method of claim 164, wherein the metabolite is selected from the group consisting of a monosaccharide, a disaccharide, a polysaccharide, a lipid, a nucleic acid, an amino acid and a polypeptide.

20

166. The method of claim 165, wherein the saccharide, disaccharide or polysaccharide comprises a glucose or a sucrose.

167. The method of claim 163, wherein the composition is selected from the group consisting of an antibiotic, a metal, a steroid and an antibody.

25

168. The method of claim 114, wherein the measured metabolic parameter comprises an increase or a decrease in the secretion of a byproduct or a secreted composition of a cell.

30

169. The method of claim 168, wherein the byproduct or secreted composition is selected from the group consisting of a toxin, a lymphokine, a polysaccharide, a lipid, a nucleic acid, an amino acid, a polypeptide and an antibody.

170. The method of claim 114, wherein the real time monitoring simultaneously measures a plurality of metabolic parameters.
- 5 171. The method of claim 170, wherein real time monitoring of a plurality of metabolic parameters comprises use of a Cell Growth Monitor device.
172. The method of claim 171, wherein the Cell Growth Monitor device is a Wedgewood Technology, Inc., Cell Growth Monitor model 652.
- 10 173. The method of claim 171, wherein the real time simultaneous monitoring measures uptake of substrates, levels of intracellular organic acids and levels of intracellular amino acids.
- 15 174. The method of claim 171, wherein the real time simultaneous monitoring measures: uptake of glucose; levels of acetate, butyrate, succinate or oxaloacetate; and, levels of intracellular natural amino acids.
- 20 175. The method of claim 171, further comprising use of a computer-implemented program to real time monitor the change in measured metabolic parameters over time.
176. The method of claim 175, wherein the computer-implemented program comprises a computer-implemented method as set forth in Figure 28.
- 25 177. The method of claim 176, wherein the computer-implemented method comprises metabolic network equations.
178. The method of claim 176, wherein the computer-implemented method comprises a pathway analysis.

179. The method of claim 176, wherein the computer-implemented program comprises a preprocessing unit to filter out the errors for the measurement before the metabolic flux analysis.

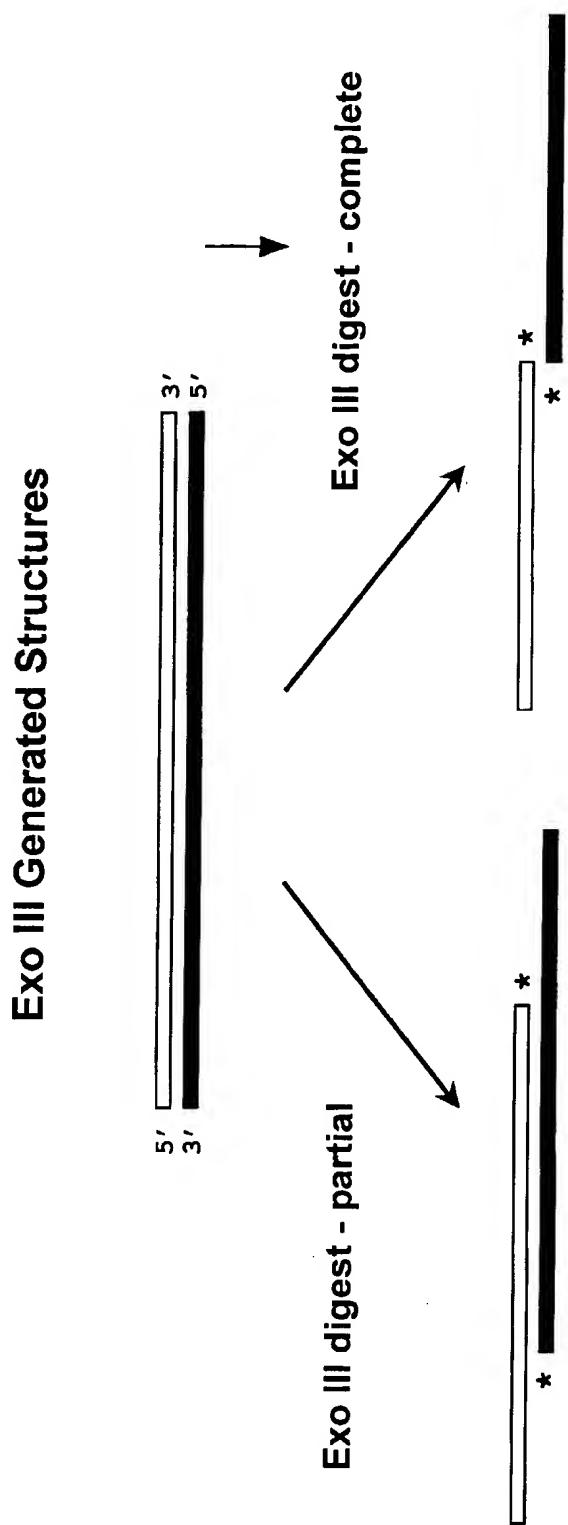


Figure 1

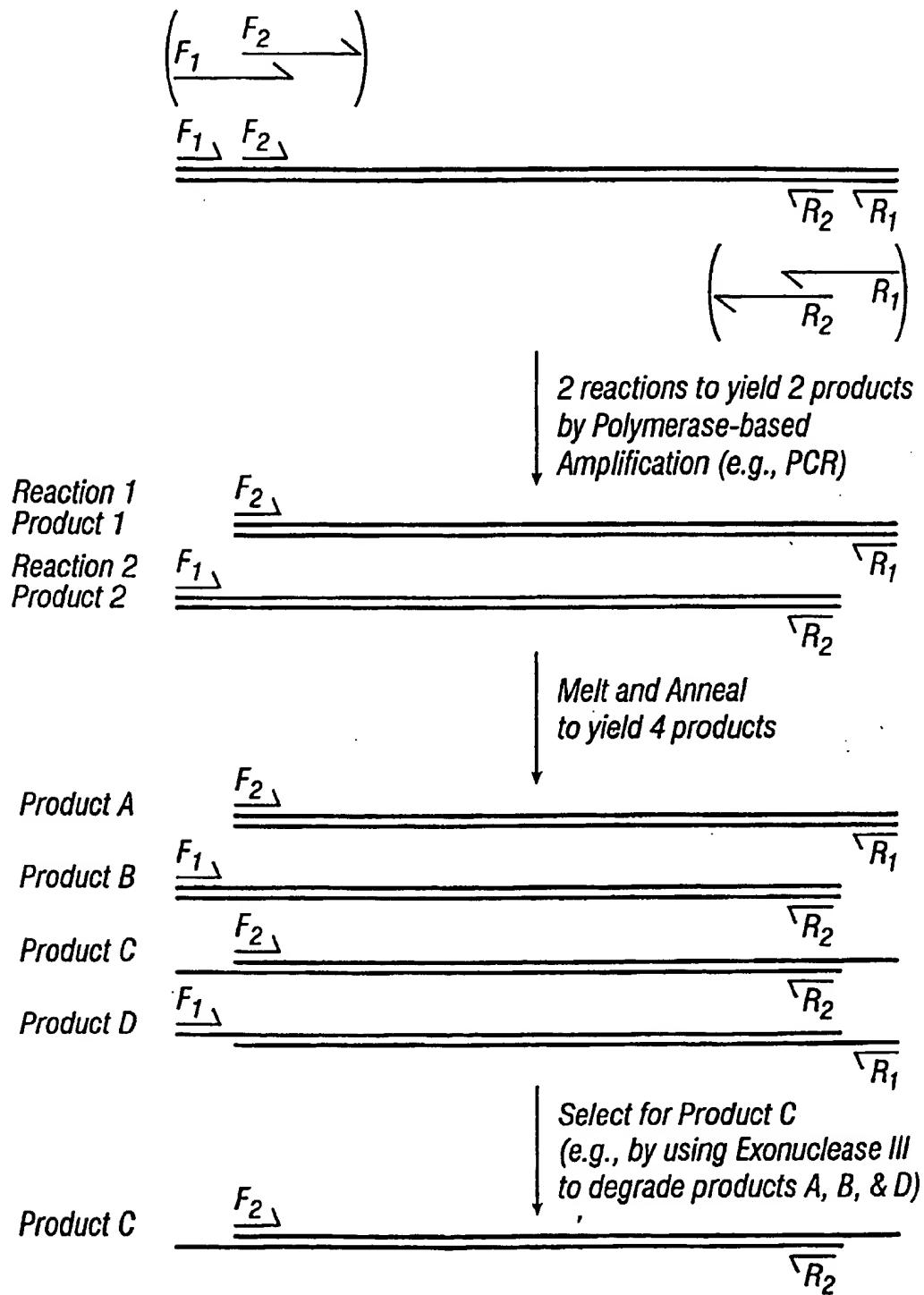
**FIG. 2**

FIGURE 3. Unique Overhangs And Unique Couplings.

The number of unique overhangs of each size (e.g. the total number of unique overhangs composed of 1 or 2 or 3, etc. nucleotides) exceeds the number of unique couplings that can result from the use of all the unique overhangs of that size. For example, the total number of unique couplings that can be made using all the 8 unique single-nucleotide 3' overhangs and single-nucleotide 5' overhangs is 4.

PANEL A. 4 unique single-nucleotide 3' overhangs are possible (i.e., A, C, G, & T). For each of these there is a complementary 3' overhang with which it can pair (i.e., T, G, C, & A, respectively), as shown.



PANEL B. However, the number of unique single-nucleotide 3' overhangs is greater than the number of unique couplings. Thus, only 2 intrinsically unique couplings exist using single-nucleotide 3' overhangs as shown.



PANEL C. Likewise, 4 unique-single nucleotide 5' overhangs are possible (i.e., A, C, G, & T). For each of these there is a complementary 5' overhang with which it can pair (i.e., T, G, C, & A, respectively), as shown.



PANEL D. However, the number of unique single-nucleotide 5' overhangs is greater than the number of unique couplings. Thus, only 2 intrinsically unique couplings exist using single-nucleotide 5' overhangs as shown.

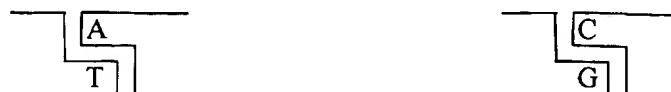
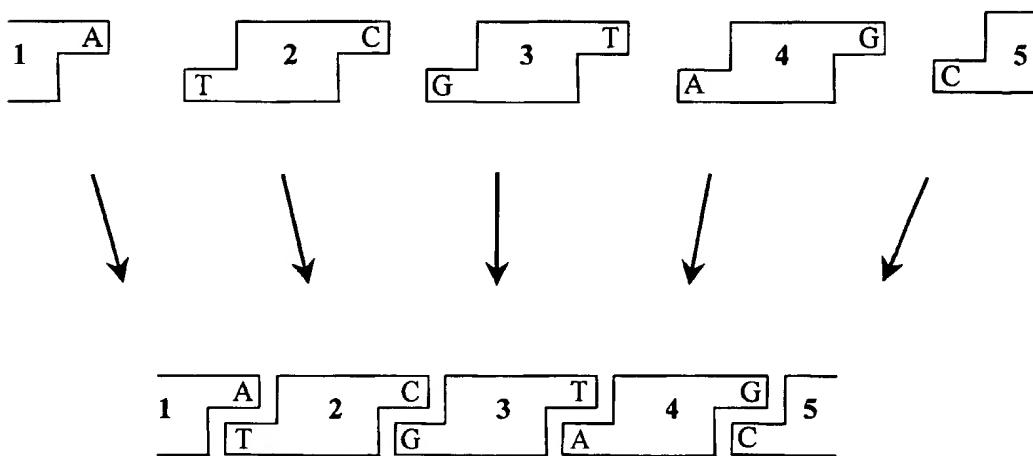


FIGURE 4. Unique Overall Assembly Order Achieved by Sequentially Coupling the Building Blocks

Awareness of the degeneracy (between the number of unique overhangs and the number of unique couplings) is important in order to avoid the production of degeneracy in the overall assembly order of the finalized nucleic acid. However, a unique overall assembly order can also be achieved - despite the use of non-unique couplings - by using building blocks having distinct combinations of couplings, and/or by stepping the assembly of the building blocks in a deliberately chosen sequence.

PANEL A. For example, one could attempt to assemble the following nucleic acid product using the 5 nucleic acid building blocks as shown.

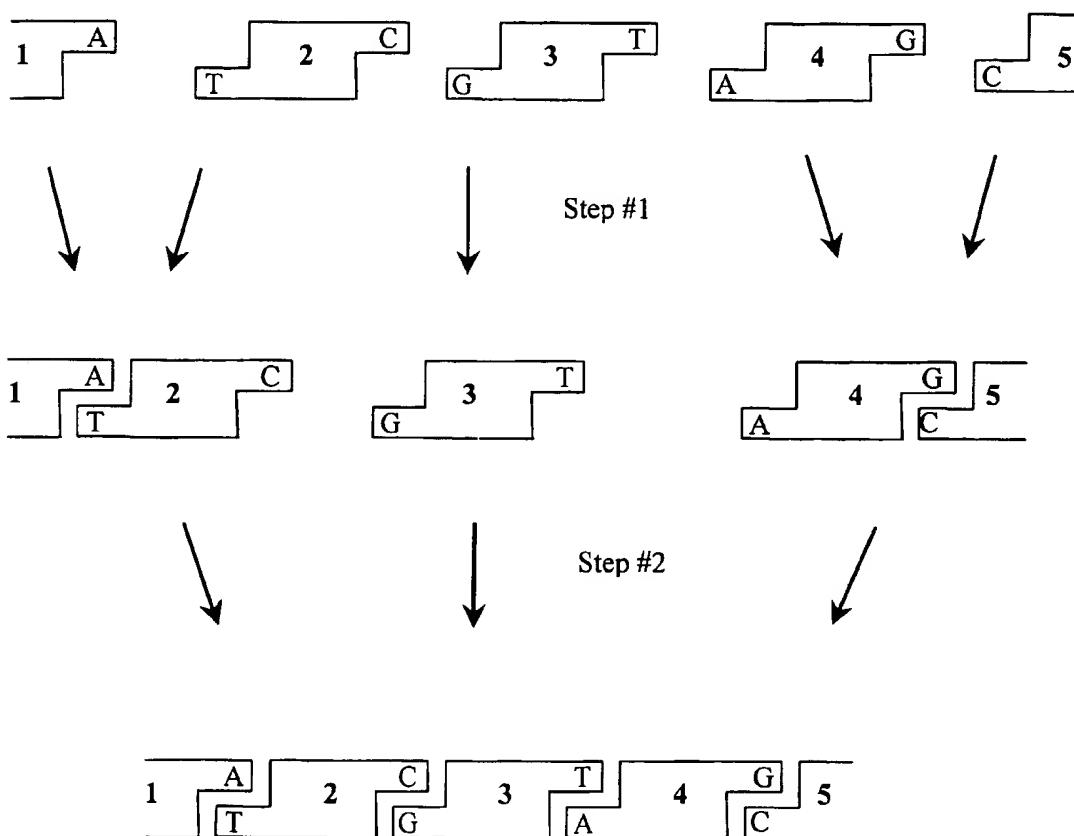


PANEL B. However, degeneracy in the overall assembly order of the 5 nucleic acid building blocks would be present if the assembly process were carried out in one step. For example, building block #2 and building block #3 could both couple to building block #1 as shown.



FIGURE 4 cont.

PANEL C. However, a unique overall assembly order could be achieved by sequentially coupling the building blocks in 2 steps (rather than all at once) as shown.



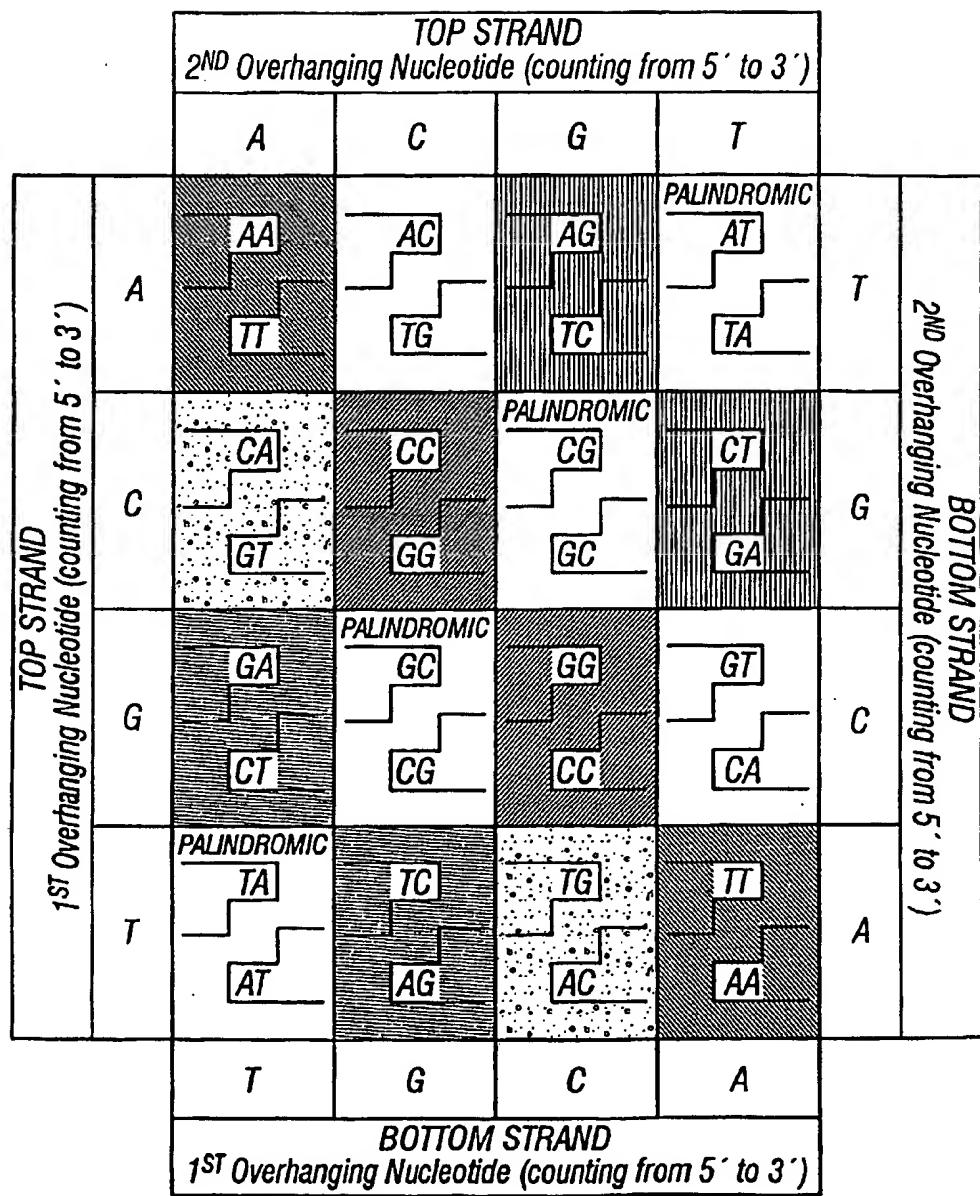


FIG. 5

Figure 6. Generation of an Exhaustive Set of Chimeric Combinations by Synthetic Ligation Reassembly.

Select for full length

Represents 15% of gene

	50	100	150
124-2d	ATGGTAAAG AACGCTTCC CGCTGTACAA CCCGGATG AGCTTATCA	TTTCACCGCA CCTTACGATCA GGTGGTATGTC	GAGACGCC AAGCCTTGT GTCGGCCACG
124-12	-ATGGTAAAG AACGCTTCC CGCTGTACAA CCCGGATG AGCTTATCA	-TTTCACCGCA CCTTACGATCA GGTGGTATGTC	-GAGACGCC AAGCCTTGT GTCGGCCACG
124-11	-ATGGTAAAG AACGCTTCC CGCTGTACAA CCCGGATG AGCTTATCA	-TTTCACCGCA CCTTACGATCA GGTGGTATGTC	-GAGACGCC AAGCCTTGT GTCGGCCACG
myo 1	CGGCGCGCG AACGCTTCC CGCTGTACAA CCCGGATG AGCTTATCA	CGGCGCGCG AACGCTTCC CGCTGTACAA CCCGGATG	CGGCGCGCG AACGCTTCC CGCTGTACAA CCCGGATG
b3	CTAACCCAAA ATGGTAAAG AACGCTTCC CGCTGTACAA CCCGGATG	CTAACCCAAA ATGGTAAAG AACGCTTCC CGCTGTACAA	CTAACCCAAA ATGGTAAAG AACGCTTCC CGCTGTACAA
b1	GCATCCGAGA ATGGTAAAG AACGCTTCC CGCTGTACAA CCCGGATG	GCATCCGAGA ATGGTAAAG AACGCTTCC CGCTGTACAA	GCATCCGAGA ATGGTAAAG AACGCTTCC CGCTGTACAA
15112	-ATGGTAAAG AACGCTTCC CGCTGTACAA CCCGGATG AGCTTATCA	-GCATCCGAGA ATGGTAAAG AACGCTTCC CGCTGTACAA	-GCATCCGAGA ATGGTAAAG AACGCTTCC CGCTGTACAA
rhod 1	CCCCCGGG AACGCTTCC CGCTGTACAA CCCGGATG AGCTTATCA	CCCCCGGG AACGCTTCC CGCTGTACAA CCCGGATG	CCCCCGGG AACGCTTCC CGCTGTACAA CCCGGATG
Com. sensus	-----G-----	T-----T-----	-GAGACGCC AAGCCTTGT GTCGGCCACG-----

+ n = 144 d.s oligos

G G G G
C C C C
CC GG
TT
A A A A
T T T T

Ligate
→

A vertical decorative element located on the right side of the page. It consists of five vertical bars of varying widths arranged in a repeating pattern. From top to bottom, there is a very narrow white bar, a wide black bar, a medium white bar, a medium black bar, and a very narrow white bar.

$8^{18} = 2 \times 10^{16}$ Reassembled Gene Variants

Figure 7. Synthetic genes from oligos.

NcoI

150am13_00	c ATGATGCACG GCGATATTTC ATCGAGCAAT GACACGGTCG GCGTTGCCGT	CCGT
150AM7_001	c ATGCATCACG GCGACATTTC ATCGAGCAAT GACACGGTCG GCGTTGCCGT	
431am7_002	c ATGAGACACG GAGATATCTC CAGCAGCAAC GATTGCGTGG GCGTGGCCGT	

150am13_00	CGTGAACTAC AAGATGCCTC GCCTTCATAC CAAGGCG GAG GT TTTAGCGA	GAG GT
150AM7_001	CGTGAACTAC AAGATGCCGC GGCTTCACAC CAAGGCT GAG GT GCTGCCA	
431am7_002	CGTGAACTAC AAGATGCCGC GGCTGCATAC CCGCCGGAG GATGGAGA	

150am13_00	ACGCCAGAAA GATCGGCGAG ATGATCGTCG GCATGAAGAC CGGCCTGCC	CGG
150AM7_001	ACTGCCGCAA GATCGCCGAC ATGCTGGTCG GCATGAAGAG CGGCCTGCC	
431am7_002	ACGCCCCCAA GATCGCCGAC ATGGTCGTGG GCATGAAGCG CGGCCTGCC	

150am13_00	GGAATGGATC TGGTGATCTT CCCGGAATAT TCGAC CCACG GCATCATGTA	CCACG
150AM7_001	GGAATGGATC TGGTGATCTT CCCGGAATAT TCCAC CCACG GCATCATGTA	
431am7_002	GGCATGGACC TGGTCATCTT CCCCGAGTAC TCCAC CCACG GCATCATGTA	

150am13_00	CGACTCCAAG GAAATGTACG ATACCGCGTC CGTCGTC CCC GCGAGGAGA	CCC GG
150AM7_001	CGACTCCAAG GAGATGTACG ACACGGCGTC GACGGTC CCG GGTGAAGAGA	
431am7_002	CGACGCCAAG GAAATGTACG AAACCGCTTC GGCCATT CCG GCGAAGAGA	

150am13_00	CCGAGATTT TGCCGAAGCC TGCCGCAAGG CGAAAGTC G GGGCGTGTTC	G GGG
150AM7_001	CCGAGATTT CGCCGAGGCC TGCCGCAAGG CCAAGGTCT G GGGCGTGTTC	
431am7_002	CTGCTGTGTT CGCCGACGCC TGCCGCAAGG CCAACGTAT G GGGCGTGTTC	

150am13_00	TCGCTCACCG GCGAACGTCA CGAGGAACAT CCGAAC AAGG CGCCCTACAA	AAAG C
150AM7_001	TCGCTGACCG GCGAGCGCCA CGAGGAGCAT CCCAA AAGG CGCCGTACAA	
431am7_002	TCGCTGACGG GCGAGCGCCA CGAAGAGCAC CGAAC AAGG CGCCGTACAA	

150am13_00	CACGCTGATC CTGATGAACG ACAAGGGCGA GGTGGTC CAG AAATACCGCA	CAG AA
150AM7_001	CACCCCTGATC CTGATGAACG ACAAGGGTGA AGTCGTT CAG AAATATCGCA	
431am7_002	CACGCTCATC CTGATGAACA ACAAGGGCGA GATCGTC CAG AAGTACCGCA	

150am13_00	AGATCATGCC GTGGGTTCCG ATCGAGGGCT GGTAT CCCC GG CAACTGCACC	GGTA
150AM7_001	AGATCATGCC GTGGGTCGCC ATCGAAGGGT GGTAT CCCC GG CAACTGCACG	
431am7_002	AGATCATGCC CTGGGTCGCC ATCGAAGGGT GGTAT CCCC GG CGATTGCACC	

150am13_00	TACGTCTCCG ACGGGCCGAA GGGCAT TGAAG GTTTCGCTGA TCATCTGC	TGAAG
150AM7_001	TACGTCTCCG AAGGCCCGAA GGGCAT TGAAG ATGTCGCTGA TCATCTGC	
431am7_002	TATGTGTCGG AAGGCCCGAA GGGACT TGAAG ATCAGCCTCA TCATCTGC	

150am13_00	TGACGGCAAC TATCCGGAAA TCTGGCGCGA CTGCGCCATG AAGGGCGCCG	TCTGGCG
150AM7_001	CGACGGCAAC TACCCGGAAA TCTGGCGTGA CTGCGCGATG AAGGGCGCCG	
431am7_002	CGACGGCAAT TACCCCGAGA TCTGGCGCGA TTGCGCCATG CGCGGCGCCG	

Figure 7 cont.

CCAG

150am13_00	AGCTGATCGT	GGCTACATGT	ATCCGGCAA	GGACCAGCAG
150AM7_001	AACTGATCAT	GGCTACATGT	ATCCCGCAA	GGATCAGCAG
431am7_002	AGCTGATCGT	GGATAACATGT	ACCCGGCAA	GGACCAGCAG

GC

150am13_00	GTCATCATGG	CGAACCGAAT	GGCGTGGCG	AATAATTGTT	ACGTCGCGGT
150AM7_001	GTGCTGATGG	CGAAAGGAAT	GGCCTGGCC	AAACAACGTTT	ATGTCGCGGT
431am7_002	GTCATGGTGT	CCAAGGCGAT	GGCGTGGATG	AAACAACGTC	ACGTGGCGGT

GGGCTTCG

150am13_00	TTCCAATGCC	GGGGCTTCG	ATGGCGTCTA	TTCGTATTT	GGCCACTCGG
150AM7_001	CGCCAATGCC	TCGGGCTTCG	ACGGCGTCTA	CTCGTATTT	GGCCATTTCG
431am7_002	GGCCAATGCC	GGGGCTTCG	ACGGCGTGT	TTCCCTACTTC	GGCCATTTCG

TTCGA

150am13_00	CGATCATCGG	CTTCGATGGC	CGCACGCTCG	GCGAATGCGG	CGAGGAAGAA
150AM7_001	CGATCATCGG	CTTCGACGGC	CGTACCCCTCG	GCGAATGCGG	CGAGGAGGAT
431am7_002	CCATCATCGG	CTTCGACGGC	CGCACGCTGG	GCGAATGCGG	TGAAGAAGAC

C AGTA

150am13_00	TACGGCATCC	AGTATGCCA	GCTTTCGAAG	ATGCTGATCC	GCGACGCCG
150AM7_001	TATGGCATCC	AGTATGCCGC	CATCTCCAAG	TCGCTGATCC	GCGACGCCG
431am7_002	ATGGGCGTGC	AGTACGCCGA	GCTCTCCACC	AGCCTGATCC	GCGACGCCG

CAATC

150am13_00	CCGCACCGGA	CAATCGAAA	ACCATCTCTT	CAAGCTGGTG	CATCGTGGCT
150AM7_001	CCGCACCGGC	CAATCGAAA	ACCATCTCTT	CAAGCTGGTG	CACCGTGGCT
431am7_002	CAAGAACATG	CAGTCGCAGA	ACCACTTGT	CAAGCTGGTG	CACCGCAGCT

GATCAA

150am13_00	ACACCGGGTT	GATCAACTCC	GGCGAGGGCG	ACCGCGGTCT	CGCGGCCTGT
150AM7_001	ACACCGGCAT	GATCAACTCC	GGCGAGGGCG	ACCGCGGTGT	CGCGGCCTTGC
431am7_002	ACACCGGCAA	GATCAACTCC	GGCGAAGAGG	ACCGGGCGT	CGCGGCATGC

TTA

150am13_00	CTTATGAGT	TCTACAACAA	ATGGATCGCC	GATCCCGAAG	GCACCCCGCGA
150AM7_001	CGTATGATT	TCTATTCGAA	ATGGATCGCC	GATCCCGAGG	GTACACCGCGA
431am7_002	CGTACAAC	TCTACGCCAA	CTGGATCAAC	GATCCGGAGG	GCACCGCGCAA

ATGGT

150am13_00	ATGGTTCGAG	TCCTTTACCC	GGCCGACGGT	GGGAACCGAT	GAAGCGCCCA
150AM7_001	GATGGTGGAA	TCCTTCACGC	GTCCGACGGT	GGGTGTGGAG	GAATGCCCGA
431am7_002	GATGGTGGAA	TCCTTCACCC	GGTCCACCGT	GGGCACGCCG	GAGTGCCCCA

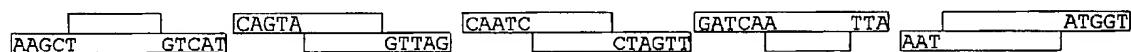
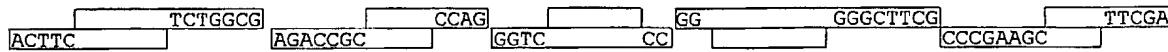
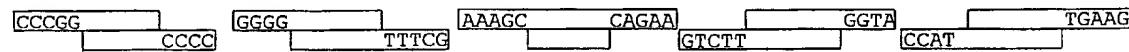
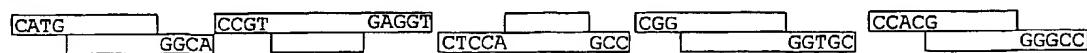
TCGAG

150am13_00	TCGAAGGCAT	CCCGAACAAAG	GTCGCGGTGC	ACCGCTGA	aagct
150AM7_001	TCGAAGGGCAT	TCCGAACAAAG	GCCACCAACGC	ACCGCTGA	aagct
431am7_002	TGGAAAGGCAT	CCCCAACGAG	GACGCCAAGC	ACCGCTAG	aagct

HindIII

Figure 8. Nucleic acid building blocks for synthetic ligation gene reassembly.

NcoI



HindIII

Figure 9. Addition of Introns by Synthetic Ligation Reassembly.

NcoI



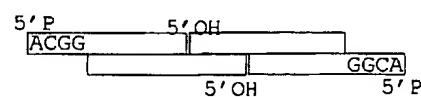
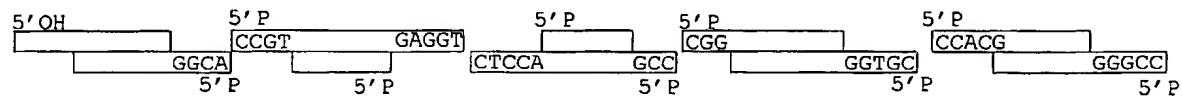
Figure 10. Ligation Reassembly Using Fewer Than All The Nucleotides Of An Overhang.

Gap Ligation



Ligation of one strand only;
gap in second strand can be repaired in vivo

Figure 11. Avoidance of unwanted self-ligation in palindromic couplings.



No self ligation of end primers
with palindromic overhangs

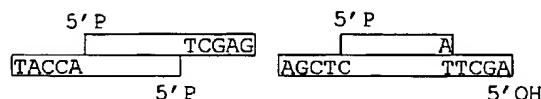


Figure 12. Pathway Engineering

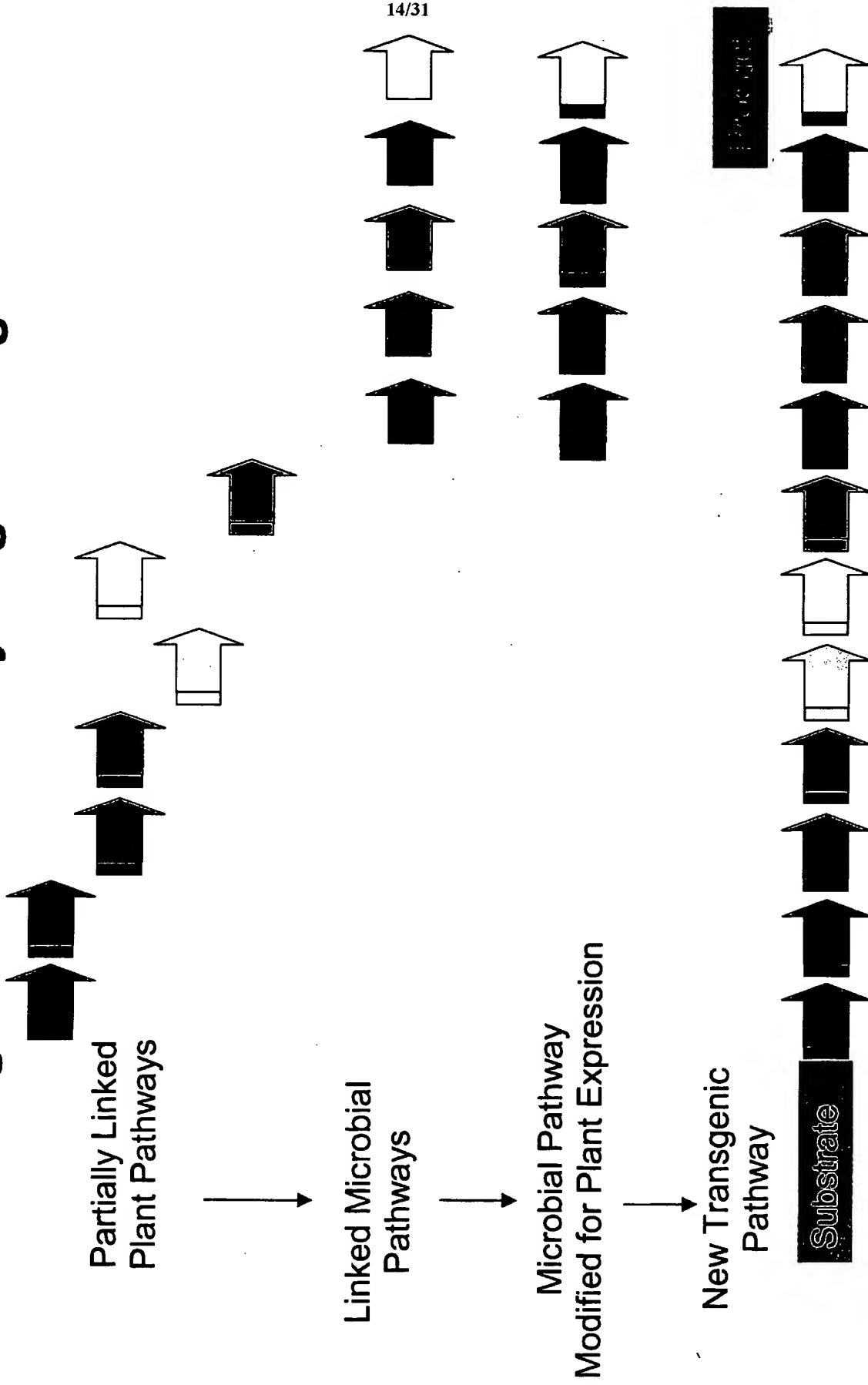


Figure 13. Pathway Improvement / Evolution

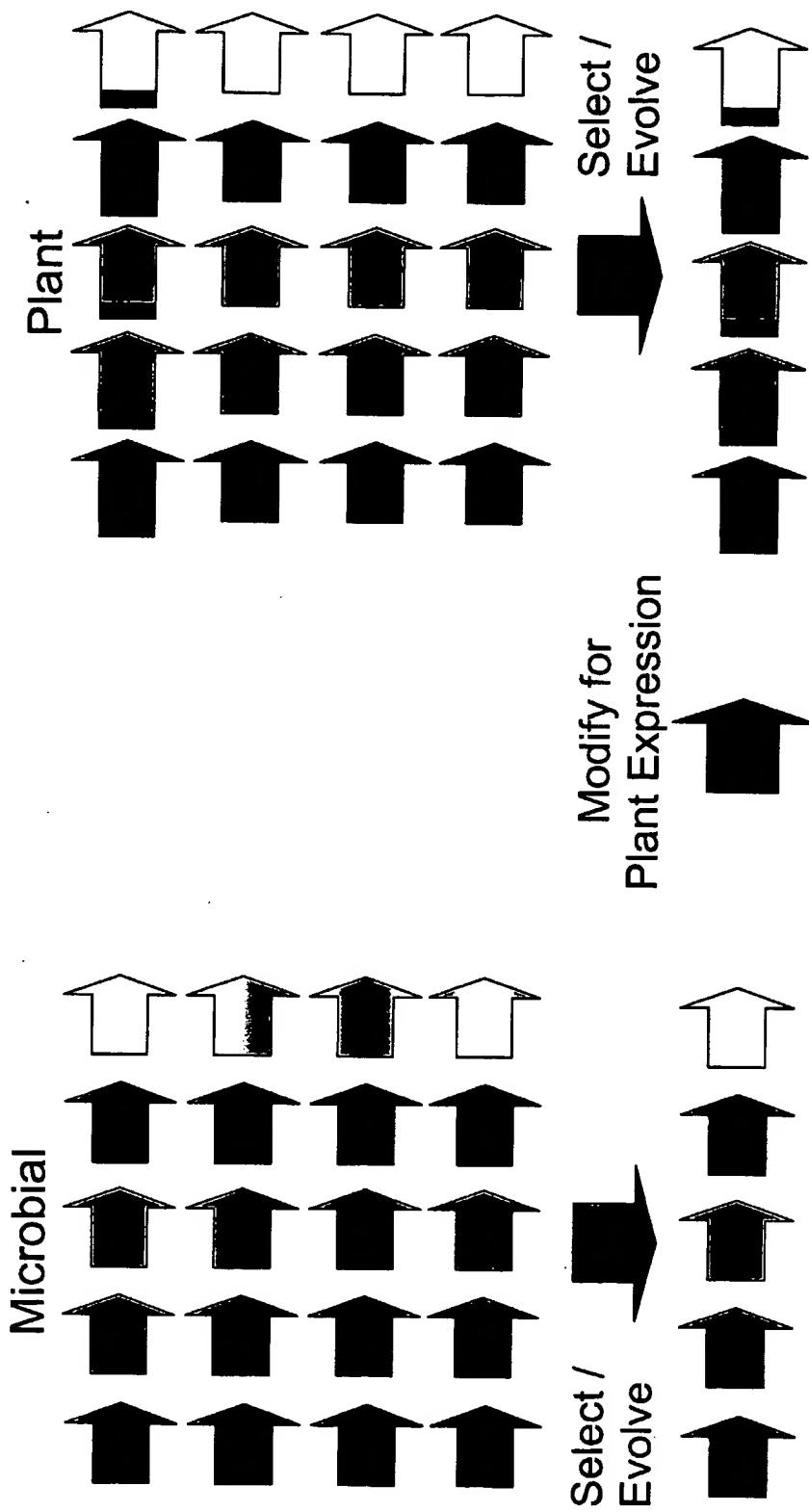


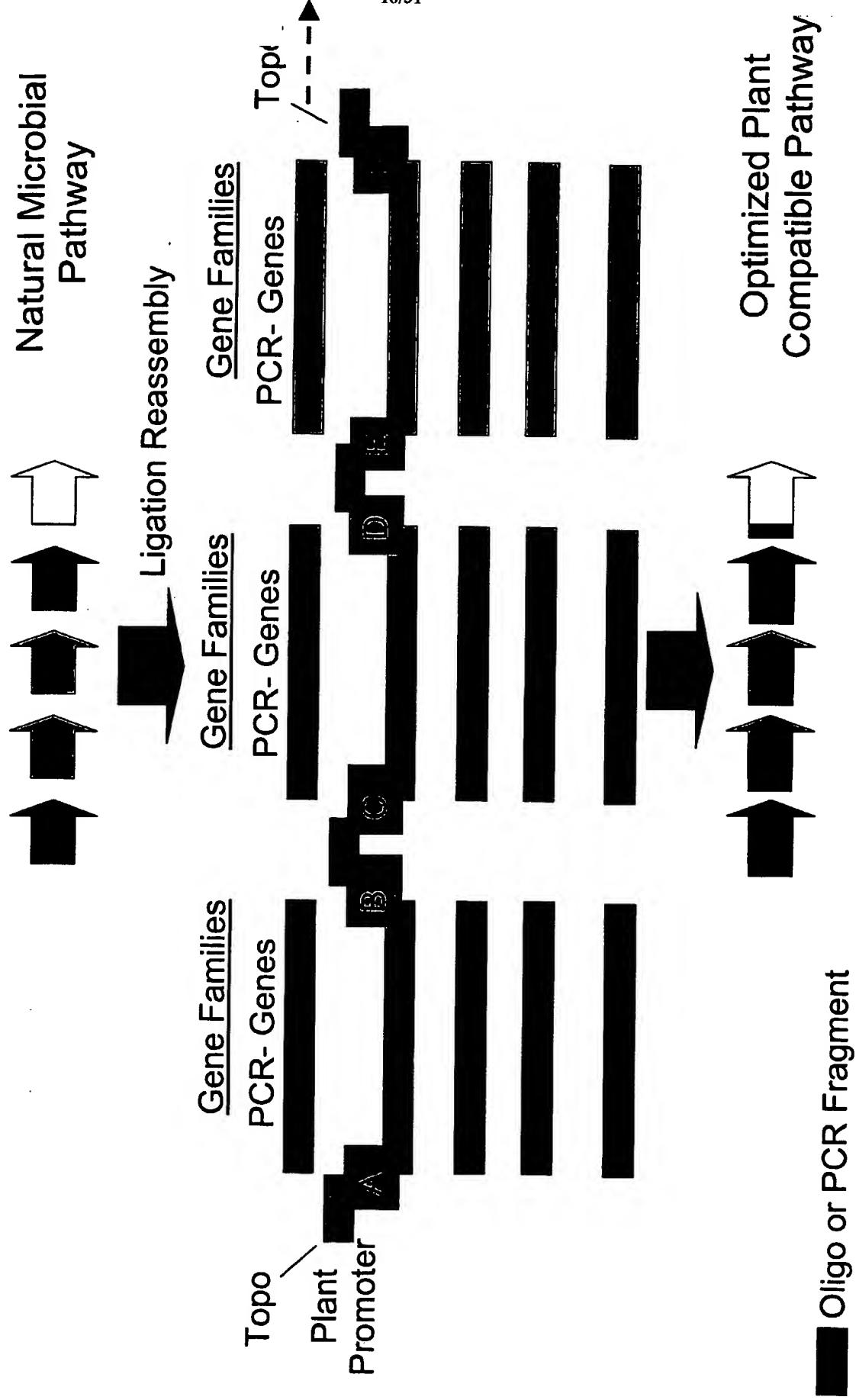
Fig 14. Conversion of Microbial Pathways to Eukaryotic Pathways

Fig. 15. ENGINEERING OF DIFFERENTIALLY ACTIVATABLE STACKED TRAITS IN NOVEL TRANSGENIC PLANTS USING DIRECTED EVOLUTION AND WHOLE CELL MONITORING

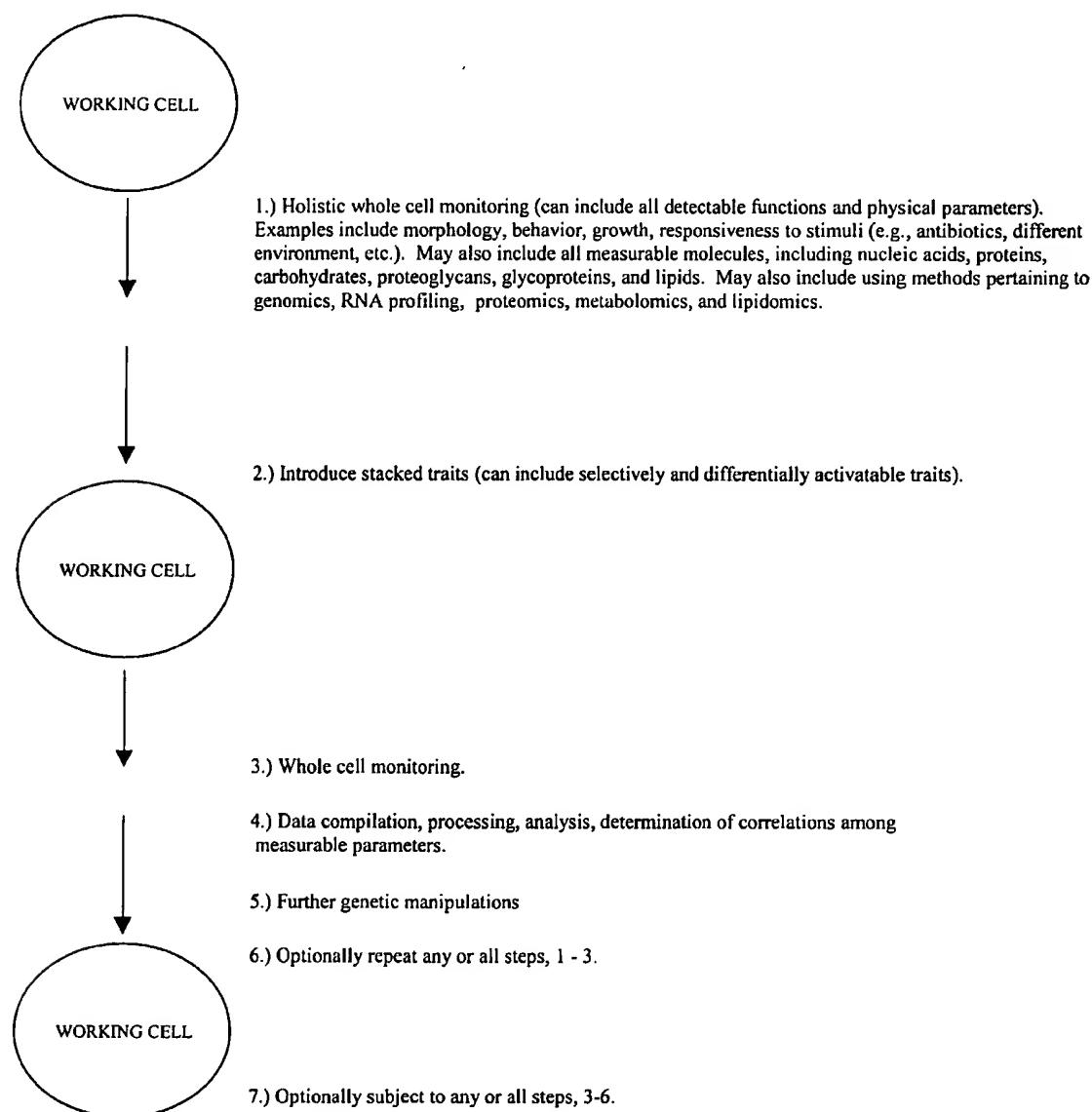


Fig. 16. Differential Activation of Selected Traits Can Be Achieved by Adjusting and Controlling the Environment of the Traits.

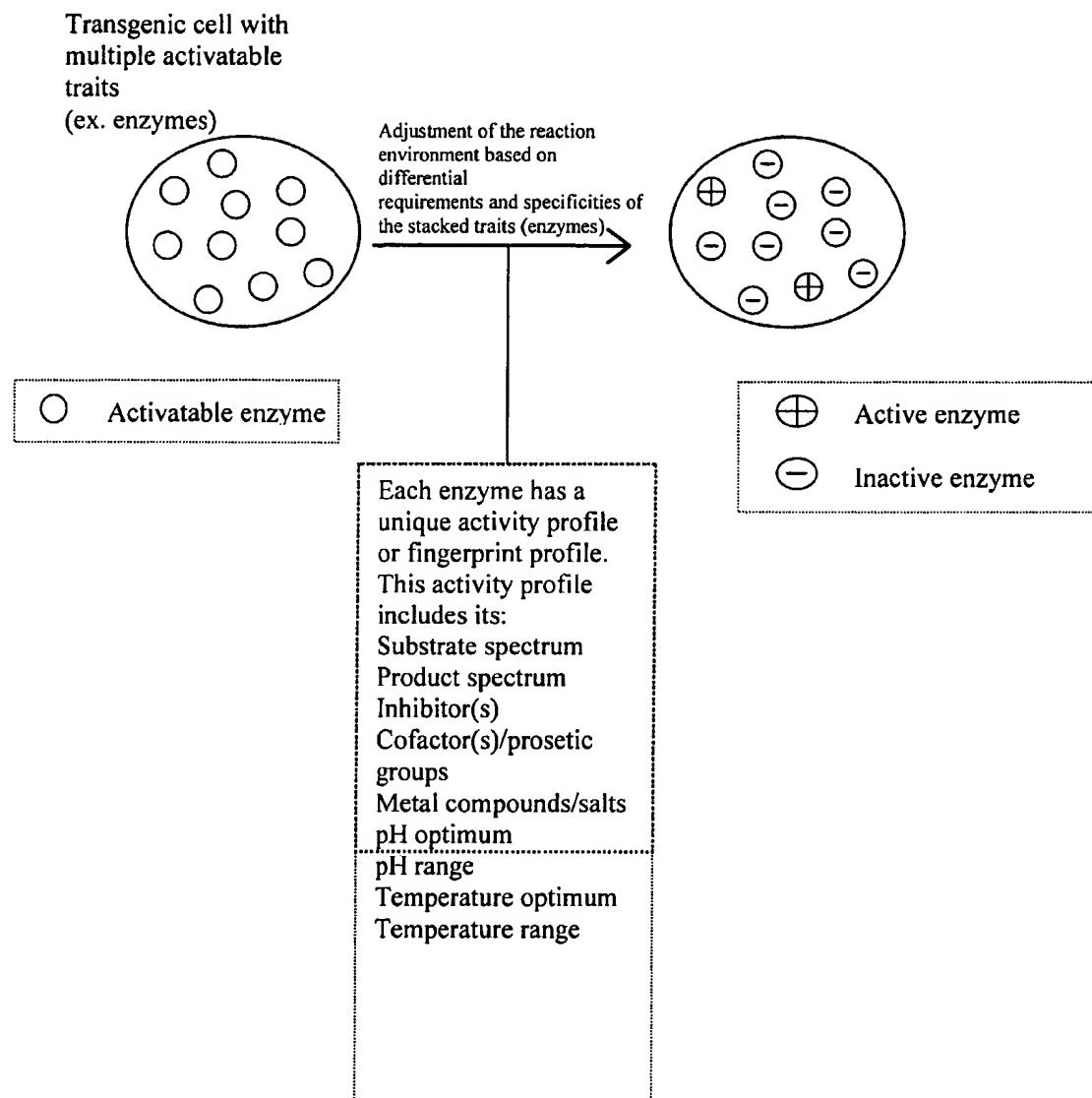


Fig. 17. Desired or improved traits for harvesting, processing, and storage conditions.

Differentially activated and/or selected enzymes respond to the environments of harvesting, processing and storage to activate environmentally-responsive specific promoters.

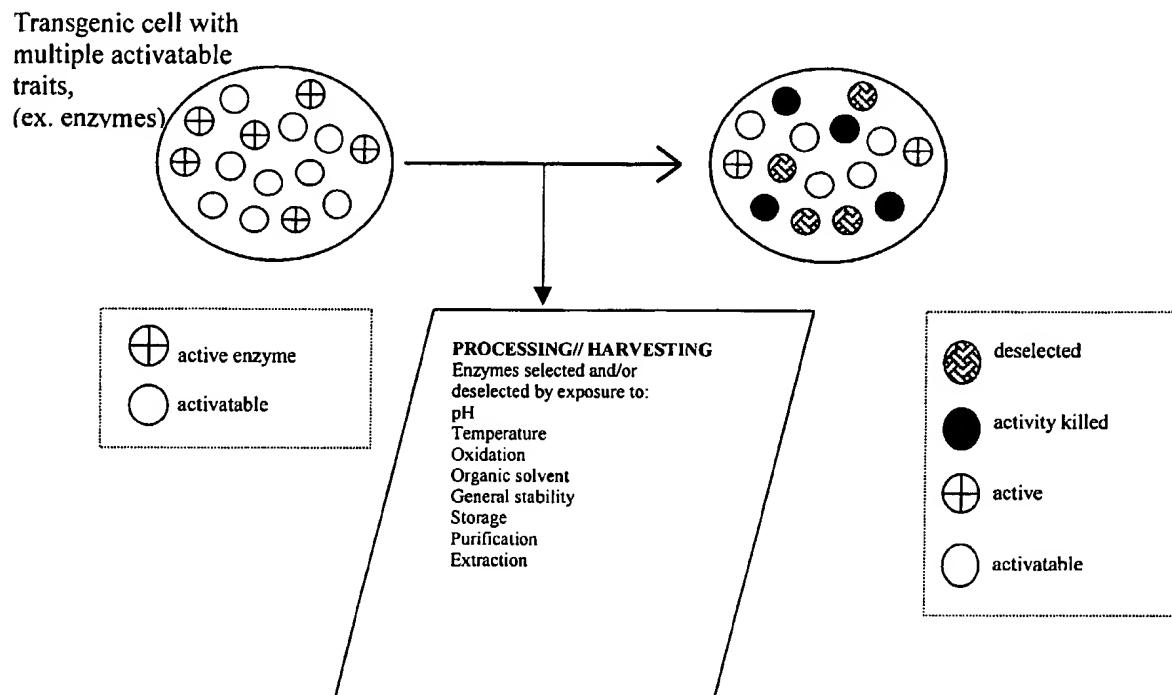


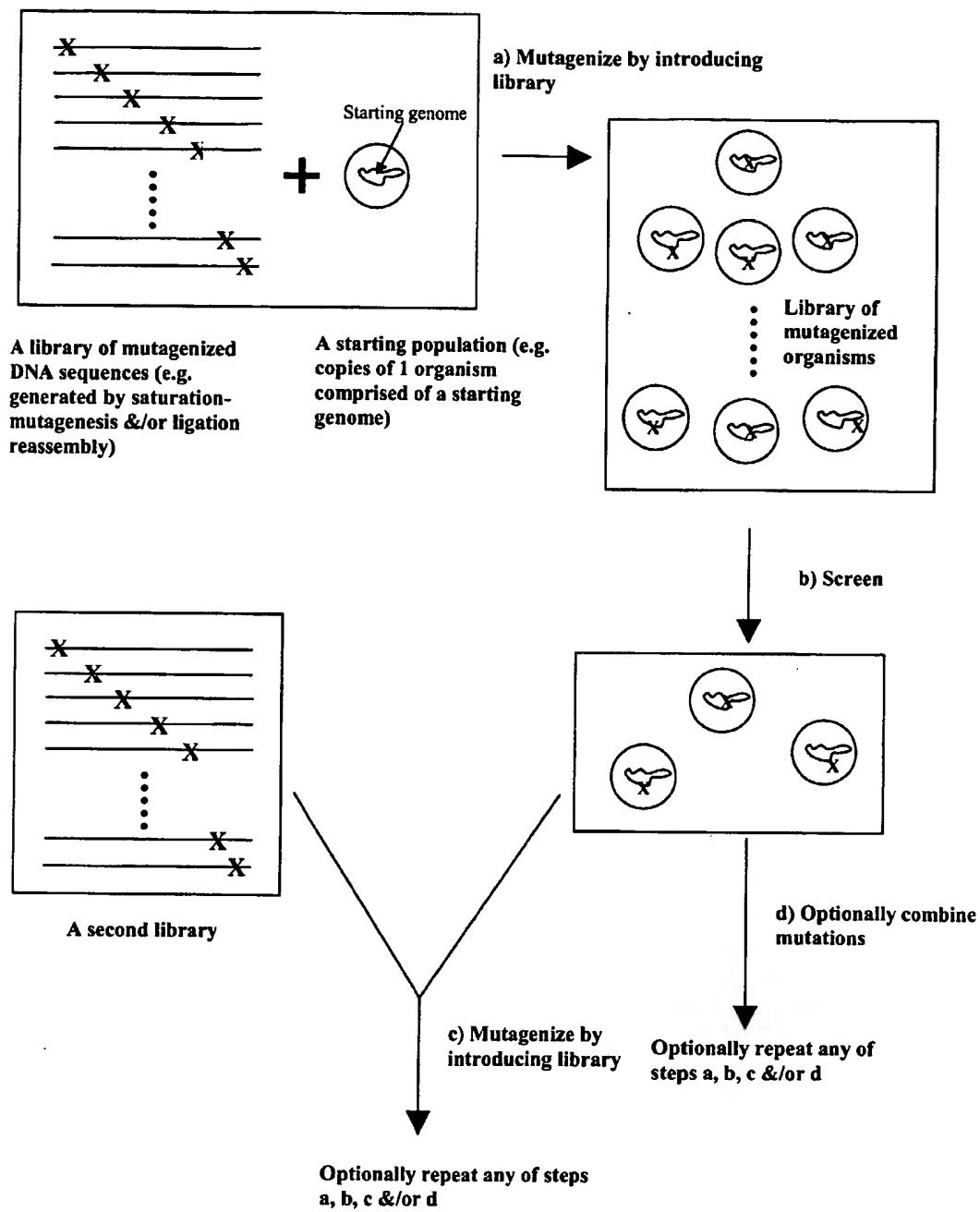
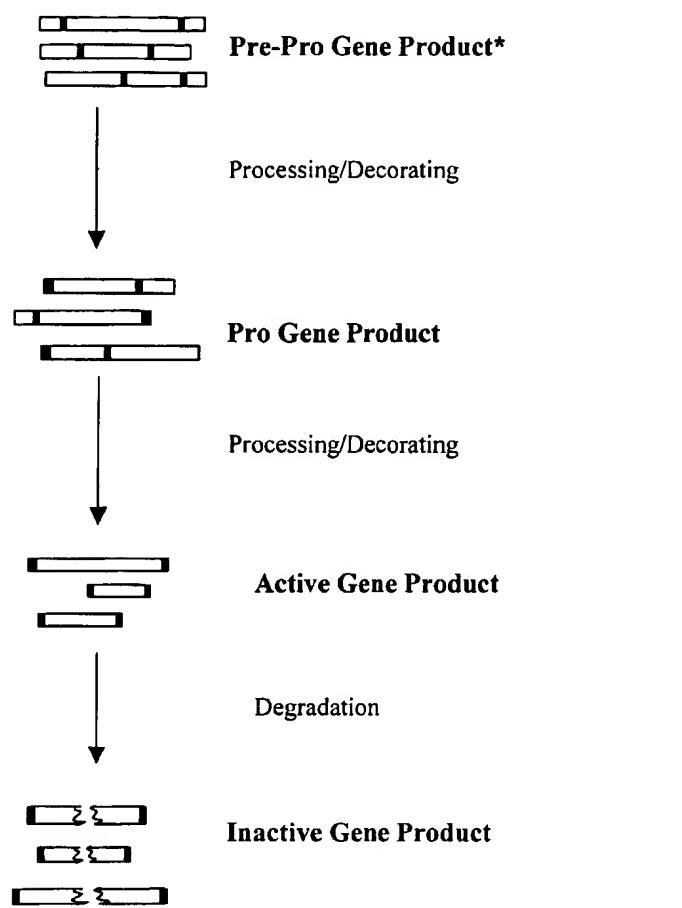
Fig. 18. Cellular Mutagenesis.

Fig. 19. Gene Product Processing

* An example of a Gene Product might be a protein. Through processing/decorating the protein changes forms, eventually becoming active. It is at this point that specific traits can be expressed differentially.

Fig. 20. Differential Activation of Selected Precursor (Inactive) Gene Products

Differential activation of selected precursor (inactive) gene products by controlling the post-translational modifications that differentially transform selected molecules from inactive precursor form to active form. Deselection of particular molecules can also be achieved by degradation (ex. By proteolytic cleavage).

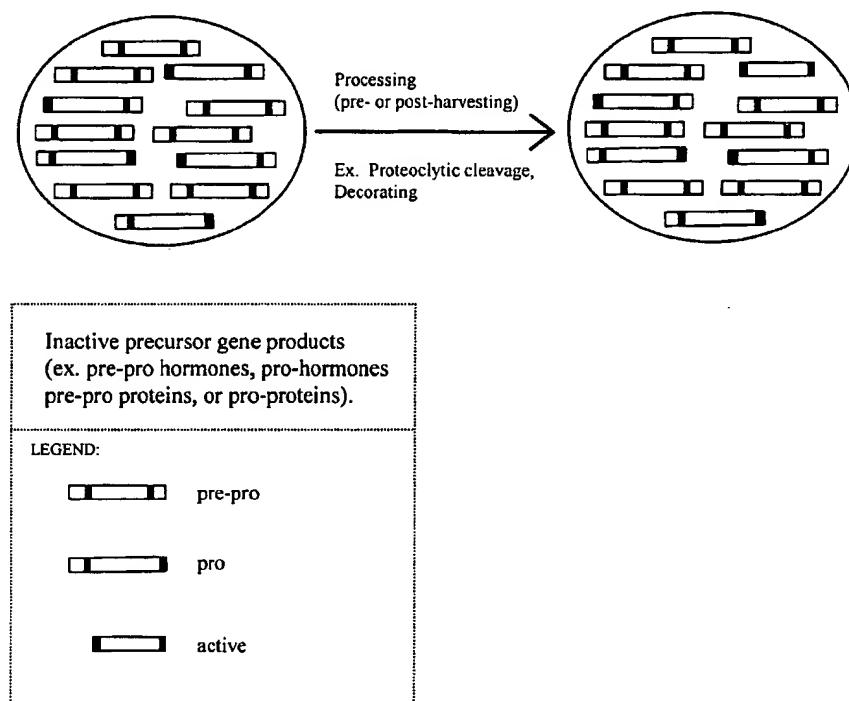


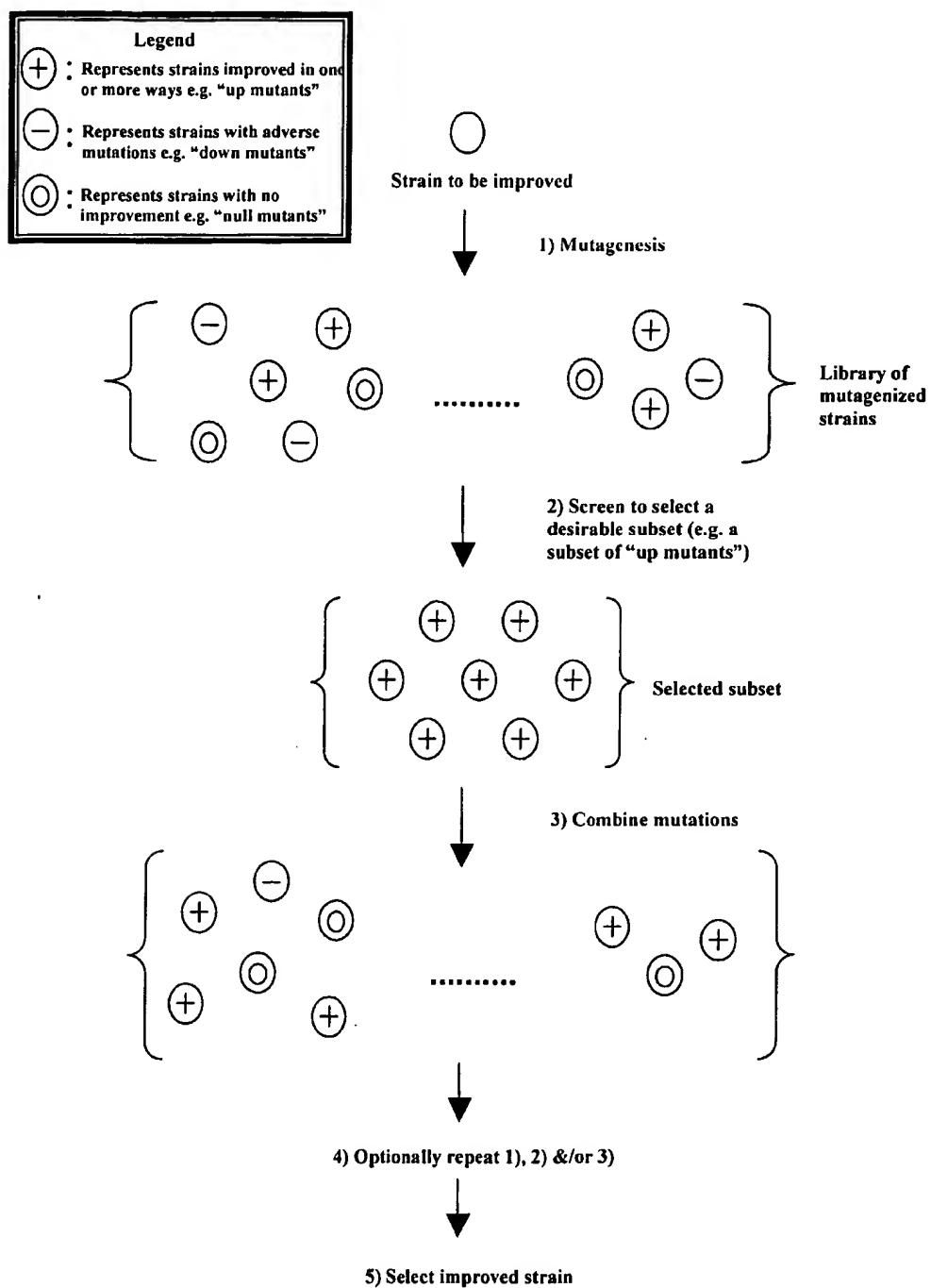
Fig. 21. Production of an improved organism or strain that has a desired trait.

Figure 22. Reassortment of polynucleotide sequences to produce an improved sequence that has a desired trait.

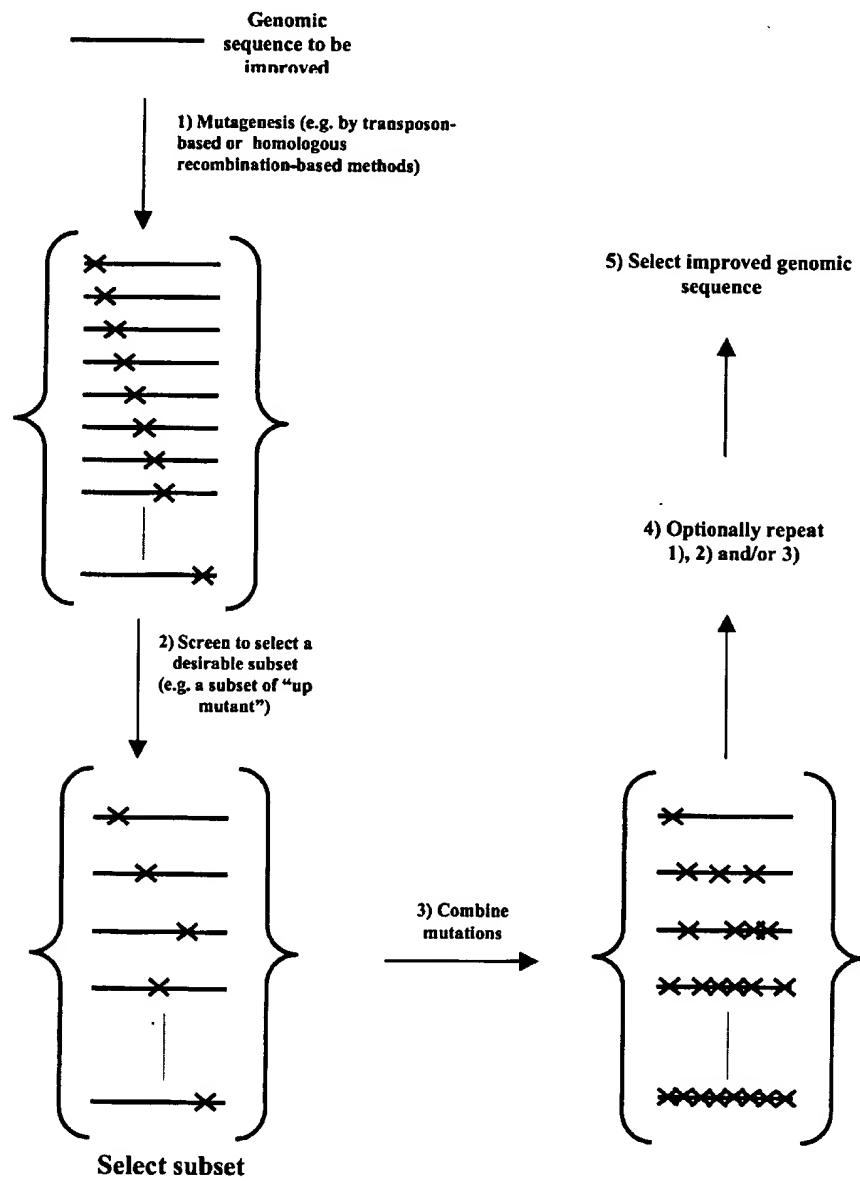


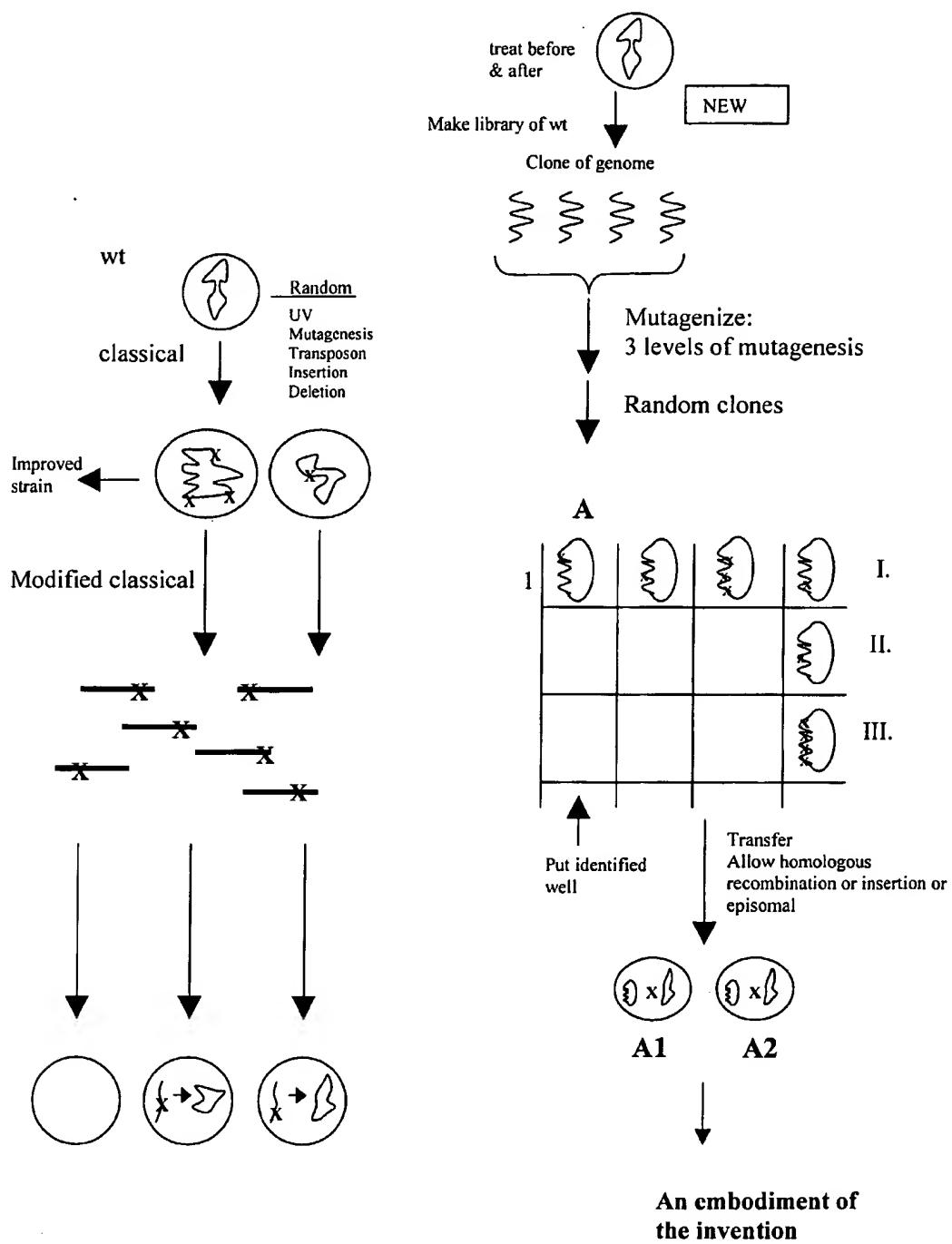
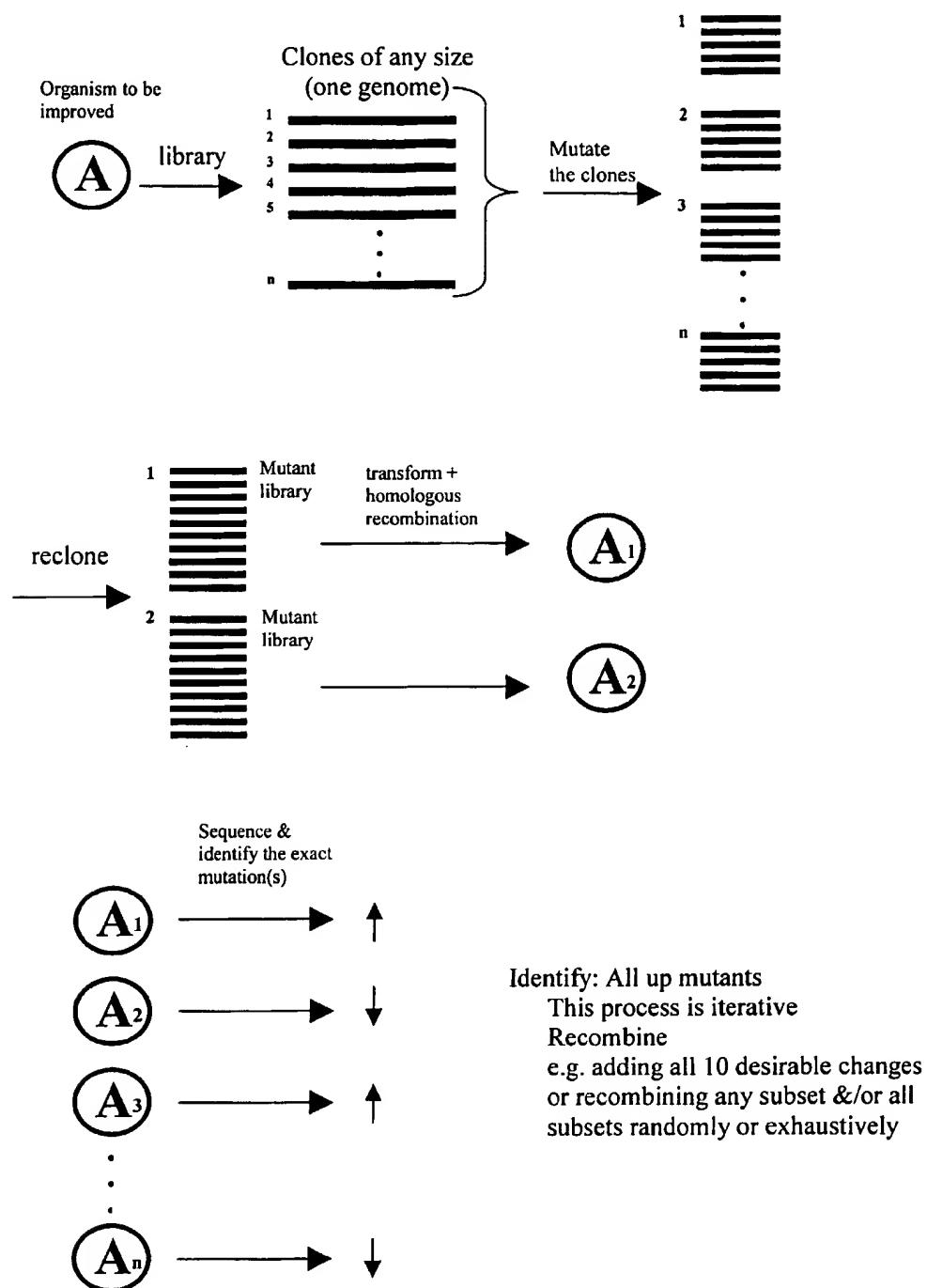
Fig. 23. Strain Improvement.

Fig. 24. Iterative Strain Improvement.

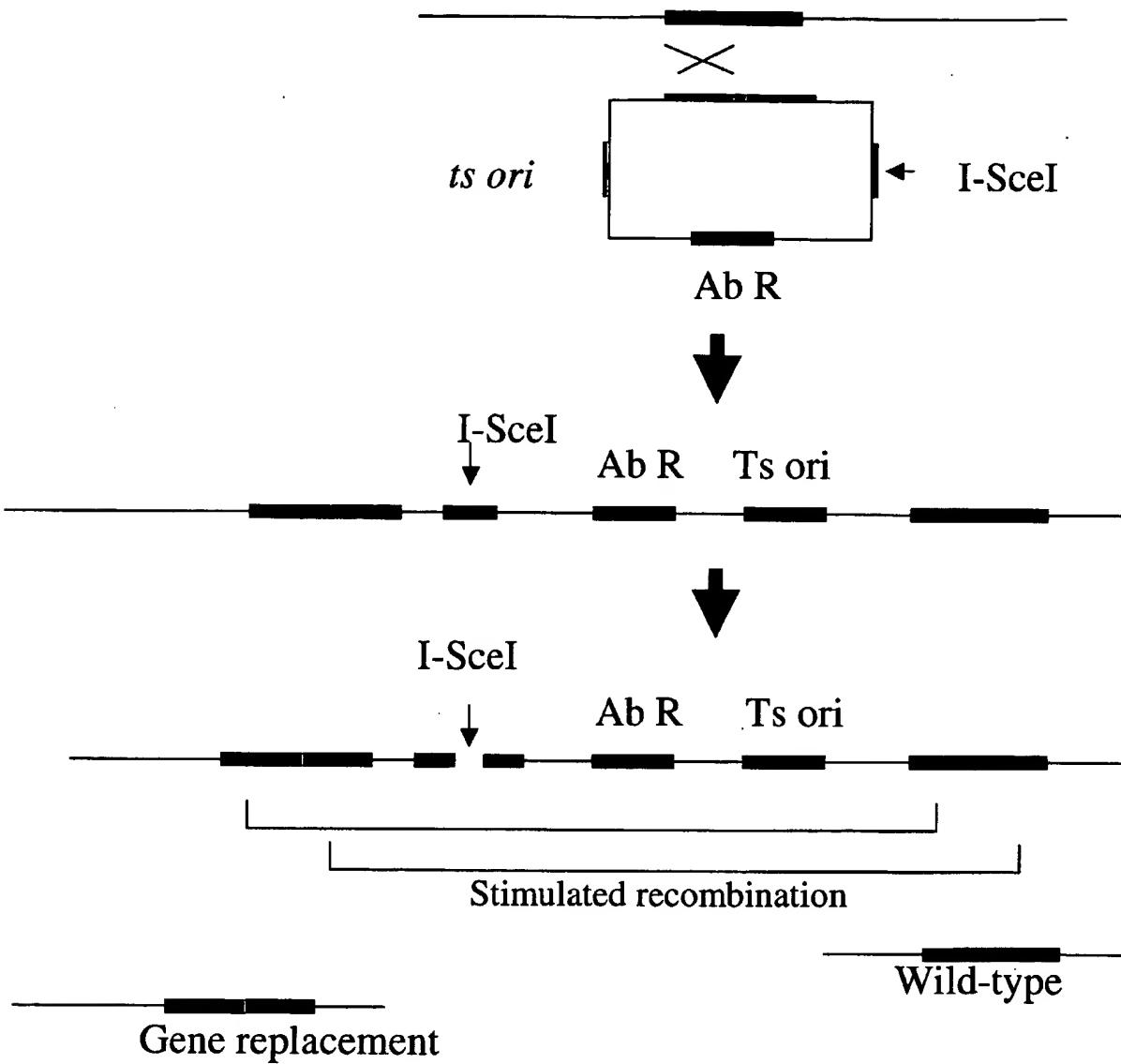
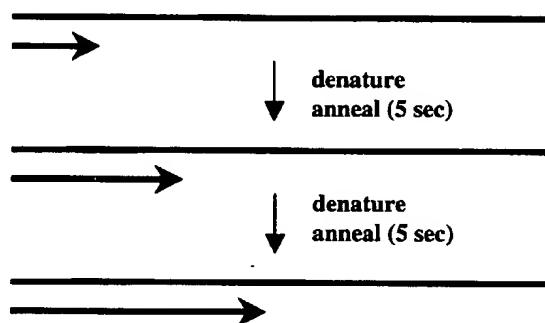
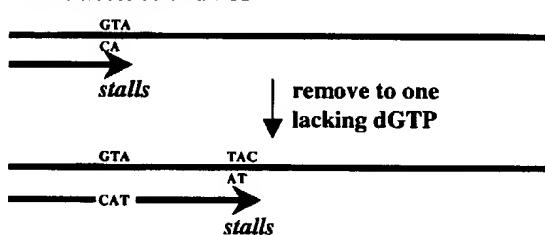
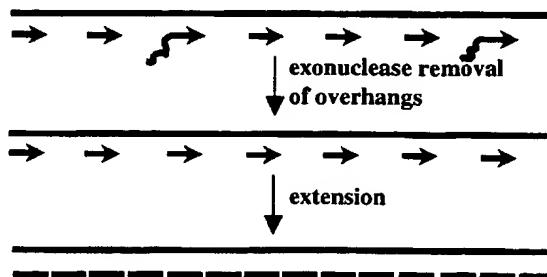
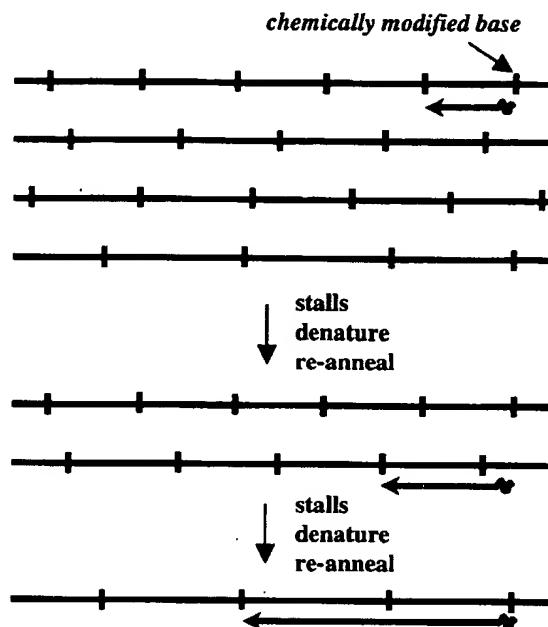


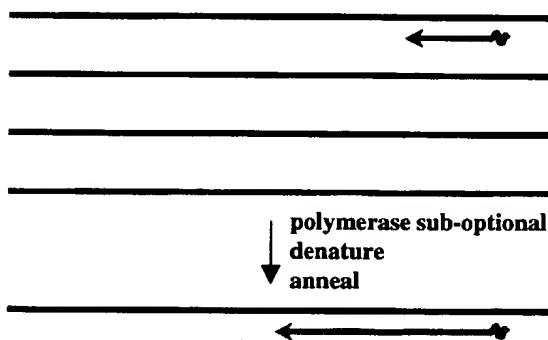
Figure 25. Illustrative diagram for the introduction of mutations for genome site saturated mutagenesis.

Fig. 26 INTERRUPTED SYNTHESIS**1) Time****2) Limited dNTP concentration****In the absence of dTTP****3) Multiple monobinders priming one polybinder template**

4) Template chemistry causes interrupted synthesis



5) Decreased DNA polymerase activity causes interrupted synthesis



6) Modified nucleotides cause interrupted synthesis

dNTP mixture: dATP, dTTP, dGTP, ddCTP

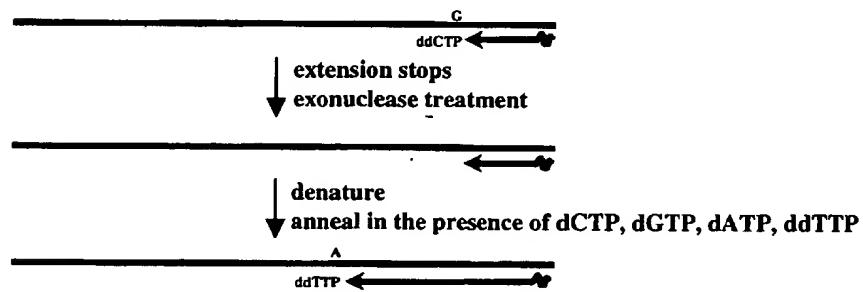


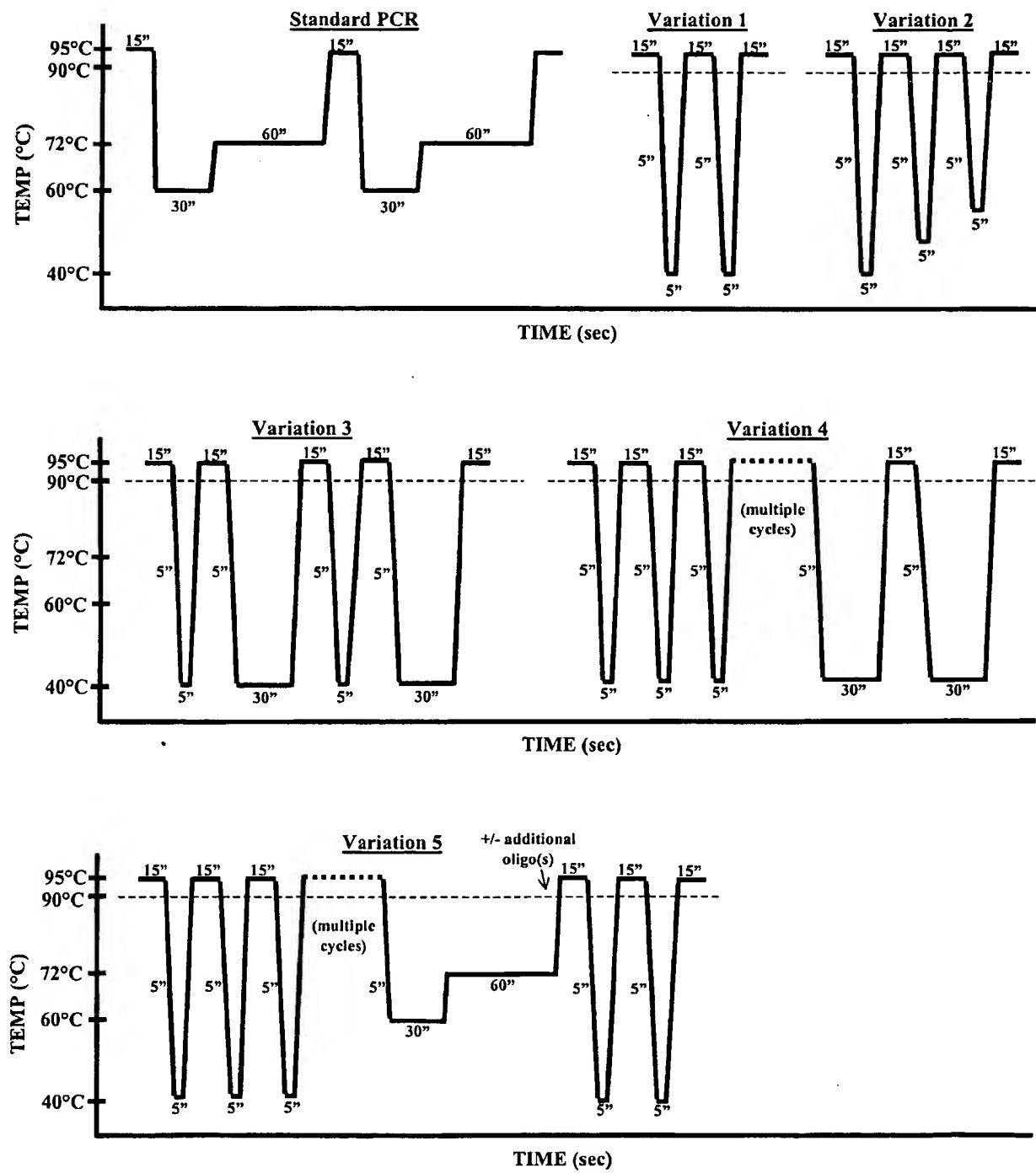
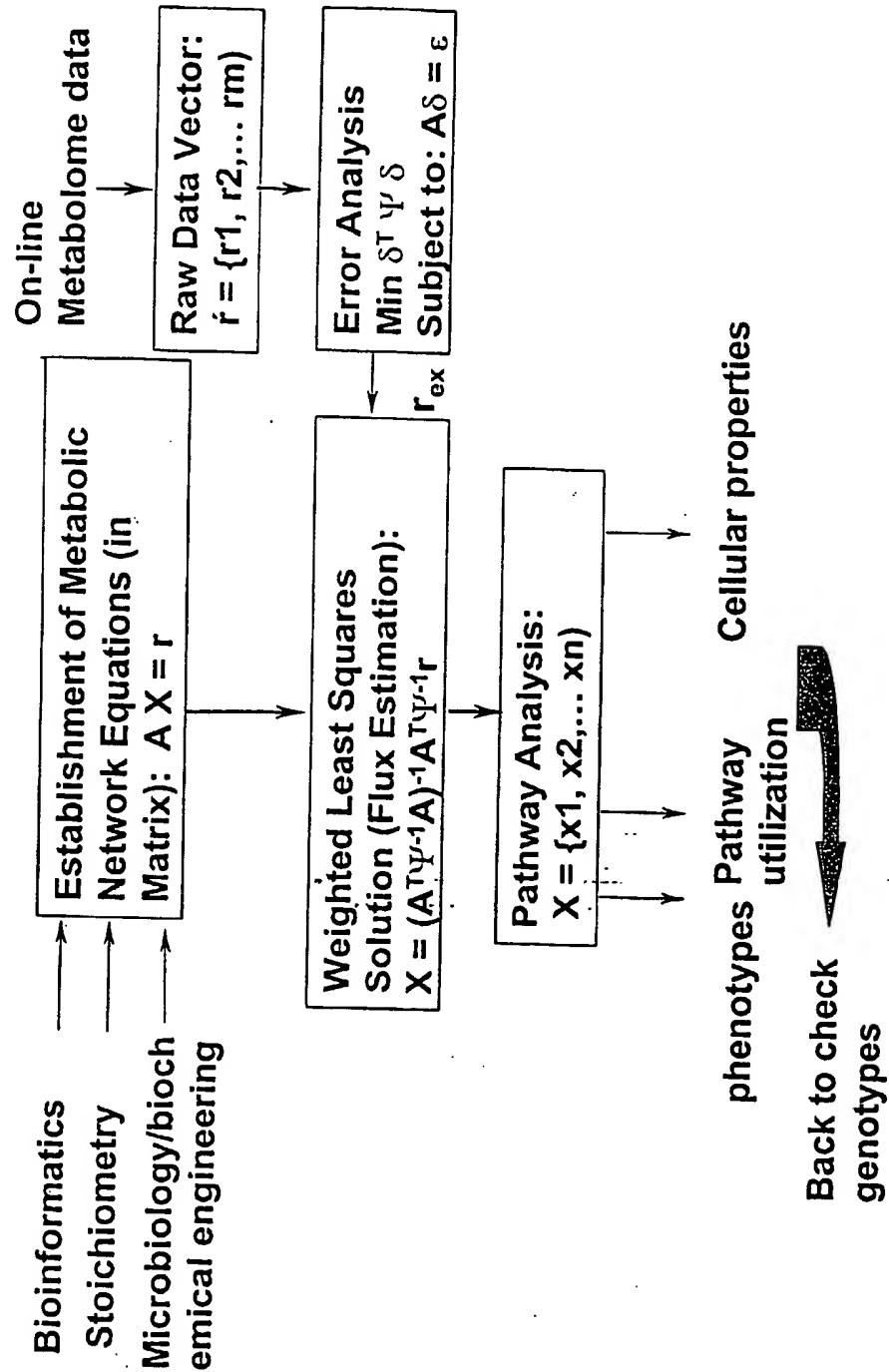
Fig. 27 INTERRUPTING SYNTHESIS BY LIMITING EXTENSION TIMES

Figure 28

Procedure for On-Line Metabolic Flux Analysis (MFA)



bacteria. The better the nitrogen fixing bacteria grow in the new host, the more copies of their recombined genes will be present for the next round of recombination. This growth rate differentiating selection is described above in detail.

4.7.8.11 BIODETECTORS / BIOSENSORS

Bioluminescence or fluorescence genes can be used as reporters by fusing them to specific regulatory genes (Cameron et. al. Appl. Biochem Biotechnol, 38:105- 140 (1993)). A specific example is one in which the luciferase genes luxCDABE of *Vibrio fischeri* were fused to the regulatory region of the isopropylbenzene catabolism operon from *Pseudomonas putida* RE204.

Transformation of this fusion construct into *E. coli* resulted in a strain which produced light in response to a variety of hydrophobic compound such as substituted benzenes, chlorinated solvents and naphthalene (Selifonova et. al., Appl Environ Microbiol 62:778-783 (1996)). This type of construct is useful for the detection of pollutant levels, and has the added benefit of only measuring those pollutants that are bioavailable (and therefore potentially toxic). other signal molecules such as jellyfish green fluorescent protein could also be fused to genetic regulatory regions that respond to chemicals in the environment. This should allow a variety of molecules to be detected by their ability to induce expression of a protein or proteins which result in light, fluorescence or some other easily detected signal. Recursive sequence recombination can be used in several ways to modify this type of biodetection system. It can be used to increase the amplitude of the response, for example by increasing the fluorescence of the green fluorescent protein. Recursive sequence recombination could also be used to increase induced expression levels or catalytic activities of other signal-generating systems, for example of the luciferase genes.

Recursive sequence recombination can also be used to alter the specificity of biosensors. The regulatory region, and transcriptional activators that interact with this region and with the chemicals that induce transcription can also be stochastic &/or

non-stochastic mutagenized. This should generate regulatory systems in which transcription is activated by analogues of the normal inducer, so that biodetectors for different chemicals can be developed.

In this case, selection would be for constructs that are activated by the (new) specific chemical to be detected. Screening could be done simply with fluorescence (or light) activated cell sorting, since the desired improvement is in light production. In addition to detection of environmental pollutants, biosensors can be developed that will respond to any chemical for which there are receptors, or for which receptors can be evolved by recursive sequence recombination, such as hormones, growth factors, metals and drugs. These receptors may be intracellular and direct activators of transcription, or they may be membrane bound receptors that activate transcription of the signal indirectly, for example by a phosphorylation cascade. They may also not act on transcription at all, but may produce a signal by some post-transcriptional modification of a component of the signal generating pathway. These receptors may also be generated by fusing domains responsible for binding different ligands with different signaling domains. Again, recursive sequence recombination can be used to increase the amplitude of the signal generated to optimize expression and functioning of chimeric receptors, and to alter the specificity of the chemicals detected by the receptor.

4.8 PROMOTING GENETIC EXCHANGE

Some methods of the invention effect recombination of cellular DNA by propagating cells under conditions inducing exchange of DNA between cells. DNA exchange can be promoted by generally applicable methods such as electroporation, biolistics, cell fusion, or in some instances, by conjugation, transduction, or agrobacterium mediated transfer and meiosis. For example, Agrobacterium can transform *S. cerevisiae* with T-DNA, which is incorporated into the yeast genome by both homologous recombination and a gap repair mechanism. (Piers et al., Proc. Natl. Acad. Sci. USA 93(4), 1613-8 (1996)).

In some methods, initial diversity between cells (i.e., before genome exchange) is induced by chemical or radiation-induced mutagenesis of a progenitor cell type, optionally followed by screening for a desired phenotype. In other methods, diversity is natural as where cells are obtained from different individuals, strains or species.

In some stochastic &/or non-stochastic mutagenesis methods, induced exchange of DNA is used as the sole means of effecting recombination in each cycle of recombination. In other methods, induced exchange is used in combination with natural sexual recombination of an organism. In other methods, induced exchange and/or natural sexual recombination are used in combination with the introduction of a fragment library. Such a fragment library can be a whole genome, a whole chromosome, a group of functionally or genetically linked genes, a plasmid, a cosmid, a mitochondrial genome, a viral genome (replicative and nonreplicative) or specific or random fragments of any of these. The DNA can be linked to a vector or can be in free form. Some vectors contain sequences promoting homologous or nonhomologous recombination with the host genome. Some fragments contain double stranded breaks such as caused by shearing with glass beads, sonication, or chemical or enzymatic fragmentation, to stimulate recombination. In each case, DNA can be exchanged between cells after which it can undergo recombination to form hybrid genomes. Generally, cells are recursively subject to recombination to increase the diversity of the population prior to screening. Cells bearing hybrid genomes, e.g., generated after at least one, and usually several cycles of recombination are screened for a desired phenotype, and cells having this phenotype are isolated. These cells can additionally form starting materials for additional cycles of recombination in a recursive recombination/selection scheme.

4.8.1 PROTOPLAST FUSION

One means of promoting exchange of DNA between cells is by fusion of cells, such as by protoplast fusion. A protoplast results from the removal from a cell of its cell wall, leaving a membrane-bound cell that depends on an isotonic or hypertonic

medium for maintaining its integrity. If the cell wall is partially removed, the resulting cell is strictly referred to as a spheroplast and if it is completely removed, as a protoplast. However, here the term protoplast includes spheroplasts unless otherwise indicated.

Protoplast fusion is described by Shaffner et al., Proc. Natl. Acad Sci. USA 77, 2163 (1980) and other exemplary procedures are described by Yoakum et al., US 4,608,339, Takahashi et al., US 4,677,066 and Sambrook et al., at Ch. 16. Protoplast fusion has been reported between strains, species, and genera (e.g., yeast and chicken erythrocyte). Protoplasts can be prepared for both bacterial and eukaryotic cells, including mammalian cells and plant cells, by several means including chemical treatment to strip cell walls. For example, cell walls can be stripped by digestion with a cell wall degrading enzyme such as lysozyme in a 10-20% sucrose, 50 mM EDTA buffer. Conversion of cells to spherical protoplasts can be monitored by phase-contrast microscopy. Protoplasts can also be prepared by propagation of cells in media supplemented with an inhibitor of cell wall synthesis, or use of mutant strains lacking capacity for cell wall formation. Preferably, eukaryotic cells are synchronized in G1 phase by arrest with inhibitors such as -factor, K. lactis killer toxin, leflonamide and adenylate cyclase inhibitors. Optionally, some but not all, protoplasts to be fused can be killed and/or have their DNA fragmented by treatment with ultraviolet irradiation, hydroxylamine or cupferon (Reeves et al., FFMS Microbiol. Lett. 99, 193 - 198 (1992)). In this situation, killed protoplasts are referred to as donors, and viable protoplasts as acceptors.

Using dead donors cells can be advantageous in subsequently recognizing fused cells with hybrid genomes, as described below. Further, breaking up DNA in donor cells is advantageous for stimulating recombination with acceptor DNA. Optionally, acceptor and/or fused cells can also be briefly, but nonlethally, exposed to UV irradiation further to stimulate recombination.

Once formed, protoplasts can be stabilized in a variety of osmolytes and compounds such as sodium chloride, potassium chloride, sodium phosphate, potassium phosphate, sucrose, sorbitol in the presence of DTT. The combination of buffer, pH, reducing agent, and osmotic stabilizer can be optimized for different cell types. Protoplasts can be induced to fuse by treatment with a chemical such as PEG, calcium chloride or calcium propionate or electrofusion (Tsoneva, Acta Microbiologica Bulgaria 24, 53-59 (1989)). A method of cell fusion employing electric fields has also been described. See Chang US, 4,970,154. Conditions can be optimized for different strains.

The fused cells are heterokaryons containing genomes from two or more component protoplasts. Fused cells can be enriched from unfused parental cells by sucrose gradient sedimentation or cell sorting. The two nuclei in the heterokaryons can fuse (karyogamy) and homologous recombination can occur between the genomes. The chromosomes can also segregate asymmetrically resulting in regenerated protoplasts that have lost or gained whole chromosomes. The frequency of recombination can be increased by treatment with ultraviolet irradiation or by use of strains overexpressing recA or other recombination genes, or the yeast rad genes, and cognate variants thereof in other species, or by the inhibition of gene products of MutS, MutL, or MutD. Overexpression can be either the result of introduction of exogenous recombination genes or the result of selecting strains, which as a result of natural variation or induced mutation, overexpress endogenous recombination genes. The fused protoplasts are propagated under conditions allowing regeneration of cell walls, recombination and segregation of recombinant genomes into progeny cells from the heterokaryon and expression of recombinant genes. This process can be reiteratively repeated to increase the diversity of any set of protoplasts or cells. After, or occasionally before or during, recovery of fused cells, the cells are screened or selected for evolution toward a desired property.

Thereafter a subsequent round of recombination can be performed by preparing protoplasts from the cells surviving selection/screening in a previous round. The

protoplasts are fused, recombination occurs in fused protoplasts, and cells are regenerated from the fused protoplasts. This process can again be reiteratively repeated to increase the diversity of the starting population. Protoplasts, regenerated or regenerating cells are subject to further selection or screening.

Subsequent rounds of recombination can be performed on a split pool basis as described above. That is, a first subpopulation of cells surviving selection/screening from a previous round are used for protoplast formation. A second subpopulation of cells surviving selection/screening from a previous round are used as a source for DNA library preparation.

The DNA library from the second subpopulation of cells is then transformed into the protoplasts from the first subpopulation. The library undergoes recombination with the genomes of the protoplasts to form recombinant genomes. This process can be repeated several times in the absence of a selection event to increase the diversity of the cell population. Cells are regenerated from protoplasts, and selection/screening is applied to regenerating or regenerated cells. In a further variation, a fresh library of nucleic acid fragments is introduced into protoplasts surviving selection/screening from a previous round.

Protoplast formation of donor and recipient strains, heterokaryon formation, karyogamy, recombination, and segregation of recombinant genomes into separate cells. Optionally, the recombinant genomes, if having a sexual cycle, can undergo further recombination with each other as a result of meiosis and mating. Recursive cycles of protoplast fusion, or recursive mating/meiosis is often used to increase the diversity of a cell population. After achieving a sufficiently diverse population via one of these forms of recombination, cells are screened or selected for a desired property. Cells surviving selection/screening can then be used as the starting materials in a further cycle of protoplasting or other recombination methods as noted herein.

4.8.2 PARASEXUAL REPRODUCTION

Parasexual reproduction provides a further means for stochastic &/or non-stochastic mutagenesis genetic material between cells. This process allows recombination of parental DNA without involvement of mating types or gametes. Parasexual fusion occurs by hyphal fusion giving rise to a common cytoplasm containing different nuclei. The two nuclei can divide independently in the resulting heterokaryon but occasionally fuse. Fusion is followed by haploidization, which can involve loss of chromosomes and mitotic crossing over between homologous chromosomes. Protoplast fusion is a form of parasexual reproduction.

4.8.3 SELECTION FOR HYBRID STRAINS

4.8.3.1 IDENTIFYING CELLS FORMED BY THE FUSION OF COMPONENTS OF PARENTAL CELLS FROM TWO OR MORE DISTINCT SUBPOPULATIONS

The invention provides selection strategies to identify cells formed by fusion of components from parental cells from two or more distinct subpopulations. Selection for hybrid cells is usually performed before selecting or screening for cells that have evolved (as a result of genetic exchange) to acquisition of a desired property. A basic premise of most such selection schemes is that two initial subpopulations have two distinct markers. Cells with hybrid genomes can thus be identified by selection for both markers.

4.8.3.2 METHOD WHERE ONE SUBPOPULATION HAS A MARKER

In one such scheme, at least one subpopulation of cells bears a selective marker attached to its cell membrane. Examples of suitable membrane markers include biotin, fluorescein and rhodamine. The markers can be linked to amide or thiol groups or through more specific derivatization chemistries, such as iodo-acetates,

iodoacetamides, maleimides.

For example, a marker can be attached as follows. Cells or protoplasts are washed with a buffer (e.g., PBS), which does not interfere with the chemical coupling of a chemically active ligand which reacts with amino groups of lysines or N-terminal amino groups of membrane proteins. The ligand is either amine reactive itself (e.g., isothiocyanates, succinimidyl esters, sulfonyl chlorides) or is activated by a heterobifunctional linker (e.g. EMCS, SLAB, SPDP, SMB) to become amine reactive. The ligand is a molecule which is easily bound by protein derivatized magnetic beads or other capturing solid supports. For example, the ligand can be succinimidyl activated biotin (Molecular Probes Inc.: B-1606, B-2603, S-1515, S-1582). This linker is reacted with amino groups of proteins residing in and on the surface of a cell. The cells are then washed to remove excess labeling agent before contacting with cells from the second subpopulation bearing a second selective marker.

The second subpopulation of cells can also bear a membrane marker, albeit a different membrane marker from the first subpopulation. Alternatively, the second subpopulation can bear a genetic marker. The genetic marker can confer a selective property such as drug resistance or a screenable property, such as expression of green fluorescent protein.

After fusion of first and second subpopulations of cells and recovery, cells are screened or selected for the presence of markers on both parental subpopulations. For example, fusants are enriched for one population by adsorption to specific beads and these are then sorted by FACS for those expressing a marker. Cells surviving both screens for both markers are those having undergone protoplast fusion, and are therefore more likely to have recombined genomes. Usually, the markers are screened or selected separately. Membrane-bound markers, such as biotin, can be screened by affinity enrichment for the cell membrane marker (e.g., by panning fused cells on an

affinity matrix). For example, for a biotin membrane label, cells can be affinity purified using streptavidin-coated magnetic beads (Dynal). These beads are washed several times to remove the non-fused host cells.

Alternatively, cells can be panned against an antibody to the membrane marker. In a further variation, if the membrane marker is fluorescent, cells bearing the marker can be identified by FACS. Screens for genetic markers depend on the nature of the markers, and include capacity to grow on drug-treated media or FACS selection for green fluorescent protein. If first and second cell populations have fluorescent markers of different wavelengths, both markers can be screened simultaneously by FACS sorting.

In a further selection scheme for hybrid cells, first and second populations of cells to be fused express different subunits of a heteromultimeric enzyme. Usually, the heteromultimeric enzyme has two different subunits, but heteromultimeric enzymes having three, four or more different subunits can be used. If an enzyme has more than two different subunits, each subunit can be expressed in a different subpopulation of cells (e.g., three subunits in three subpopulations), or more than one subunit can be expressed in the same subpopulation of cells (e.g., one subunit in one subpopulation, two subunits in a second subpopulation). In the case where more than two subunits are used, selection for the poolwise recombination of more than two protoplasts can be achieved.

Hybrid cells representing a combination of genomes of first, second or more subpopulation component cells can then be recognized by an assay for intact enzyme. Such an assay can be a binding assay, but is more typically a functional assay (e.g., capacity to metabolize a substrate of the enzyme). Enzymatic activity can be detected for example by processing of a substrate to a product with a fluorescent or otherwise easily detectable absorbance or emission spectrum. The individual subunits of a

heteromultimeric enzyme used in such an assay preferably have no enzymic activity in dissociated form, or at least have significantly less activity in dissociated form than associated form. Preferably, the cells used for fusion lack an endogenous form of the heteromultimeric enzyme, or at least have significantly less endogenous activity than results from heteromultimeric enzyme formed by fusion of cells.

Penicillin acylase enzymes, cephalosporin acylase and penicillin acyltransferase are examples of suitable heteromultimeric enzymes. These enzymes are encoded by a single gene, which is translated as a proenzyme and cleaved by posttranslational autocatalytic proteolysis to remove a spacer endopeptide and generate two subunits, which associate to form the active heterodimeric enzyme. Neither subunit is active in the absence of the other subunit. However, activity can be reconstituted if these separated gene portions are expressed in the same cell by co-transformation. Other enzymes that can be used have subunits that are encoded by distinct genes (e.g., faoA and faoB genes encode 3-oxoacyl-CoA thiolase of *Pseudouomonas fragi* (*Biochem. J* 328, 815-820 (1997)).

An exemplary enzyme is penicillin G acylase from *Escherichia coli*, which has two subunits encoded by a single gene. Fragments of the gene encoding the two subunits operably linked to appropriate expression regulation sequences are transfected into first and second subpopulations of cells, which lack endogenous penicillin acylase activity. A cell formed by fusion of component cells from the first and second subpopulations expresses the two subunits, which assemble to form functional enzyme, e.g., penicillin acylase. Fused cells can then be selected on agar plates containing penicillin G, which is degraded by penicillin acylase.

In another variation, fused cells are identified by complementation of auxotrophic mutants. Parental subpopulations of cells can be selected for known auxotrophic mutations. Alternatively, auxotrophic mutations in a starting population of cells can

be generated spontaneously by exposure to a mutagenic agent. Cells with auxotrophic mutations are selected by replica plating on minimal and complete media. Lesions resulting in auxotrophy are expected to be scattered throughout the genome, in genes for amino acid, nucleotide, and vitamin biosynthetic pathways. After fusion of parental cells, cells resulting from fusion can be identified by their capacity to grow on minimal media. These cells can then be screened or selected for evolution toward a desired property. Further steps of mutagenesis generating fresh auxotrophic mutations can be incorporated in subsequent cycles of recombination and screening/selection.

In variations of the above method, de novo generation of auxotrophic mutations in each round of stochastic &/or non-stochastic mutagenesis can be avoided by reusing the same auxotrophs. For example, auxotrophs can be generated by transposon mutagenesis using a transposon bearing selective marker. Auxotrophs are identified by a screen such as replica plating. Auxotrophs are pooled, and a generalized transducing phage lysate is prepared by growth of phage on a population of auxotrophic cells. A separate population of auxtrophic cells is subjected to genetic exchange, and complementation is used to selected cells that have undergone genetic exchange and recombination. These cells are then screened or selected for acquisition of a desired property. Cells surviving screening or selection then have auxotrophic markers regenerated by introduction of the transducing transposon library. The newly generated auxotrophic cells can then be subject to further genetic exchange and screening/selection.

In a further variation, auxotrophic mutations are generated by homologous recombination with a targeting vector comprising a selective marker flanked by regions of homology with a biosynthetic region of the genome of cells to be evolved. Recombination between the vector and the genome inserts the positive selection marker into the genome causing an auxotrophic mutation. The vector is in linear form before introduction of cells.

Optionally, the frequency of introduction of the vector can be increased by capping its ends with self-complementarity oligonucleotides annealed in a hair pin formation. Genetic exchange and screening/selection proceed as described above. In each round, targeting vectors are reintroduced regenerating the same population of auxotrophic markers.

In another variation, fused cells are identified by screening for a genomic marker present on one subpopulation of parental cells and an episomal marker present on a second subpopulation of cells. For example, a first subpopulation of yeast containing mitochondria can be used to complement a second subpopulation of yeast having a petite phenotype (i.e., lacking mitochondria).

In a further variation, genetic exchange is performed between two subpopulations of cells, one of which is dead. Cells are preferably killed by brief exposure to DNA fragmenting agents such as hydroxylamine, cupferon, or irradiation. Viable cells are then screened for a marker present on the dead parental subpopulation.

4.8.4 LIPOSOME MEDIATED TRANSFERS

4.8.4.1 NUCLEIC ACID FRAGMENT LIBRARIES ARE INTRODUCED INTO PROTOPLASTS

4.8.4.1.1 THE NUCLEIC ACIDS ARE ENCAPSULATED IN LIPOSOMES TO HELP UPTAKE BY PROTOPLASTS

In the methods noted above, in which nucleic acid fragment libraries are introduced into protoplasts, the nucleic acids are sometimes encapsulated in liposomes to facilitate uptake by protoplasts. Liposome-mediated uptake of DNA by protoplasts is described in Redford et al., Mol. Gen. Genet. 184, 567-569 (1981). Liposomes can efficiently deliver large volumes of DNA to protoplasts (see Deshayes et al., FMBO J. 4, 2731-2737 (1985)).

See also, Philippot and Schuber (eds) (1995) Liposomes as Tools in Basic Research

and Industry CRC press, Boca Raton, e.g., Chapter 9, Remy et al. "Gene Transfer with Cationic Amphiphiles." Further, the DNA can be delivered as linear fragments, which are often more recombinogenic than whole genomes. In some methods, fragments are mutated prior to encapsulation in liposomes. In some methods, fragments are combined with RecA and homologs, or nucleases (e.g., restriction endonucleases) before encapsulation in liposomes to promote recombination. Alternatively, protoplasts can be treated with lethal doses of nicking reagents and then fused. Cells which survive are those which are repaired by recombination with other genomic fragments, thereby providing a selection mechanism to select for recombinant (and therefore desirably diverse) protoplasts.

4.9 SHUFFLING USING FILAMENTOUS FUNGI

Filamentous fungi are particularly suited to performing the stochastic &/or non-stochastic mutagenesis methods described above. Filamentous fungi are divided into four main classifications based on their structures for sexual reproduction: Phycomycetes, Ascomycetes, Basidiomycetes and the Fungi Imperfecti. Phycomycetes (e.g., Rhizopus, Mucor) form sexual spores in sporangium.

The spores can be uni or multinucleate and often lack septated hyphae (coenocytic).

Ascomycetes (e.g., Aspergillus, Neurospora, Penicillium) produce sexual spores in an ascus as a result of meiotic division. Ascii typically contain 4 meiotic products, but some contain 8 as a result of additional mitotic division. Basidiomycetes include mushrooms, and smuts and form sexual spores on the surface of a basidium. In holobasidiomycetes, such as mushrooms, the basidium is undivided. In hemibasidiomycetes, such as ruts (Uredinales) and smut fungi (Ustilaginales), the basidium is divided. Fungi imperfecti, which include most human pathogens, have no known sexual stage.

Fungi can reproduce by asexual, sexual or parasexual means. Asexual reproduction,

involves vegetative growth of mycelia, nuclear division and cell division without involvement of gametes and without nuclear fusion. Cell division can occur by sporulation, budding or fragmentation of hyphae.

4.9.1 EVOLVE FUNGI FROM STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS TO BECOME USEFUL HOSTS FOR GENETIC ENGINEERING OF UNRELATED GENES

4.9.2 TO IMPROVE THE CAPACITY OF FUNGI TO MAKE SPECIFIC COMPOUNDS

One general goal of stochastic &/or non-stochastic mutagenesis is to evolve fungi to become useful hosts for genetic engineering, in particular for the stochastic &/or non-stochastic mutagenesis of unrelated genes. *A. nidulans* and *neurospora* are generally the fungal organisms of choice to serve as a hosts for such manipulations because of their sexual cycles and well-established use in classical and molecular genetics. Another general goal is to improve the capacity of fungi to make specific compounds (e.g. antibacterials (penicillins, cephalosporins), antifungals (e.g. echinocandins, aureobasidins), and wood-degrading enzymes). There is some overlap between these general goals, and thus, some desired properties are useful for achieving both goals.

4.9.3 MUTATOR STRAIN

Another desired property is the production of a mutator strain of fungi. Such a fungus can be produced by stochastic &/or non-stochastic mutagenesis a fungal strain containing a marker gene with one or more mutations that impair or prevent expression of a functional product. Shufflants are propagated under conditions that select for expression of the positive marker (while allowing a small amount of residual growth without expression). Shufflants growing fastest are selected to form the starting materials for the next round of stochastic &/or non-stochastic mutagenesis.

4.9.4 EXPANDED HOST RANGE SO ABLE TO FORM HETEROKARYONS WITH MORE STRAINS

Another desired property is to expand the host range of a fungus so it can form heterokaryons with fungi from other vegetative compatibility groups. Incompatibility between species results from the interactions of specific alleles at different incompatibility loci (such as the "het" loci). If two strains undergo hyphal anastomosis, a lethal cytoplasmic incompatibility reaction may occur if the strains differ at these loci. Strains must carry identical loci to be entirely compatible. Several of these loci have been identified in various species, and the incompatibility effect is somewhat additive (hence, "partial incompatibility" can occur). Some tolerant and het-negative mutants have been described for these organisms (e.g. Dales & Croft, J Gen. Microbiol. 136, 1717-1724 (1990)). Further, a tolerance gene (*tol*) has been reported, which suppresses mating-type heterokaryon incompatibility. Stochastic &/or non-stochastic mutagenesis is performed between protoplasts of strains from different incompatibility groups. A preferred format uses a five acceptor strain and a UV-irradiated dead acceptor strain. The UV irradiation serves to introduce mutations into DNA inactivating het genes. The two strains should bear different genetic markers. Protoplasts of the strain are fused, cells are regenerated and screened for complementation of markers. Subsequent rounds of stochastic &/or non-stochastic mutagenesis and selection can be performed in the same manner by fusing the cells surviving screening with protoplasts of a fresh population of donor cells. Similar to other procedures noted herein, the cells resulting from regeneration of the protoplasts are optionally refused by protoplasting and regenerated into cells one or more times prior to any selection step to increase the diversity of the resulting population of cells to be screened.

4.9.5 ABILITY TO OUTBREED WITHOUT SELF-BREEDING

Another desired property is the introduction of multiple-allelomorph heterothallism into Ascomyceles and Fungi imperfecti, which do not normally exhibit this property. This mating system allows outbreeding without self-breeding. Such a mating system

can be introduced by stochastic &/or non-stochastic mutagenesis Ascomycetes and Fungi imperfecti with DNA from Gasteromycetes or Hymenomycetes, which have such a system.

4.9.6 SPONTANEOUS FORMATION OF PROTOPLASTS

Another desired property is spontaneous formation of protoplasts to facilitate use of a fungal strain as a stochastic &/or non-stochastic mutagenesis host. Here, the fungus to be evolved is typically mutagenized. Spores of the fungus to be evolved are briefly treated with a cell-wall degrading agent for a time insufficient for complete protoplast formation, and are mixed with protoplasts from other strain(s) of fungi. Protoplasts formed by fusion of the two different subpopulations are identified by genetic or other selection/or screening as described above. These protoplasts are used to regenerate mycelia and then spores, which form the starting material for the next round of stochastic &/or non-stochastic mutagenesis. In the next round, at least some of the surviving spores are treated with cell-wall removing enzyme but for a shorter time than the previous round. After treatment, the partially stripped cells are labeled with a first label. These cells are then mixed with protoplasts, which may derive from other cells surviving selection in a previous round, or from a fresh strain of fungi. These protoplasts are physically labeled with a second label. After incubating the cells under conditions for protoplast fusion fusants with both labels are selected.

These fusants are used to generate mycelia and spores for the next round of stochastic &/or non-stochastic mutagenesis, and so forth. Eventually, progeny that spontaneously form protoplasts (i.e., without addition of cell wall degrading agent) are identified. As with other procedures noted herein, cells or protoplasts can be reiteratively fused and regenerated prior to performing any selection step to increase the diversity of the resulting cells or protoplasts to be screened. Similarly, selected cells or protoplasts can be reiteratively fused and regenerated for one or several cycles without imposing selection on the resulting cellular or protoplast populations, thereby increasing the diversity of cells or protoplasts which are eventually screened. This

process of performing multiple cycles of recombination interspersed with selection steps can be reiteratively repeated as desired.

4.9.7 ACQUISITION OR IMPROVEMENT OF GENES ENCODING IN BIOSYNTHETIC PATHWAYS; TRANSPORTER PROTEINS; AND METABOLIC FLUX

Another desired property is the acquisition and/or improvement of genes encoding enzymes in biosynthetic pathways, genes encoding transporter proteins, and genes encoding proteins involved in metabolic flux control. In this situation, genes of the pathway can be introduced into the fungus to be evolved either by genetic exchange with another strain of fungus possessing the pathway or by introduction of a fragment library from an organism possessing the pathway. Genetic material of these fungi can then be subjected to further stochastic &/or non-stochastic mutagenesis and screening/selection by the various procedures discussed in this application.

Shufflant strains of fungi are selected/screened for production of the compound produced by the metabolic pathway or precursors thereof.

4.9.8 INCREASED STABILITY TO EXTREME CONDITIONS

Another desired property is increasing the stability of fungi to extreme conditions such as heat. In this situation, genes conferring stability can be acquired by exchanging DNA with or transforming DNA from a strain that already has such properties.

Alternatively, the strain to be evolved can be subjected to random mutagenesis. Genetic material of the fungus to be evolved can be stochastic &/or non-stochastic mutagenized by any of the procedures described in this application, with shufflants being selected by surviving exposure to extreme conditions.

4.9.9 GROWTH UNDER ALTERED NUTRITIONAL REQUIREMENTS

Another desired property is capacity of a fungus to grow under altered nutritional requirements (e.g., growth on particular carbon or nitrogen sources). Altering nutritional requirements is particularly valuable, e.g., for natural isolates of fungi that produce valuable commercial products but have esoteric and therefore expensive nutritional requirement. The strain to be evolved undergoes genetic exchange and/or transformation with DNA from a strain that has the desired nutritional requirements. The fungus to be evolved can then optionally be subjected to further stochastic &/or non-stochastic mutagenesis as described in this application and with recombinant strains being selected for capacity to grow in the desired nutritional circumstances. Optionally, the nutritional circumstances can be varied in successive rounds of stochastic &/or non-stochastic mutagenesis starting at close to the natural requirements of the fungus to be evolved and in subsequent rounds approaching the desired nutritional requirements.

4.9.10 NATURAL COMPETANCE TO TAKE UP A PLASMID BEARING A SELECTIVE MARKER

Another desired property is acquisition of natural competence in a fungus. The procedure for acquisition of natural competence by stochastic &/or non-stochastic mutagenesis is generally described in PCT/US97/04494. The fungus to be evolved typically undergoes genetic exchange or transformation with DNA from a bacterial strain or fungal strain that already has this property.

Cells with recombinant genomes are then selected by capacity to take up a plasmid bearing a selective marker. Further rounds of recombination and selection can be performed using any of the procedures described above.

4.9.11 REDUCED OR INCREASED SECRETION OF PROTEASES AND DNASES

Another desired property is reduced or increased secretion of proteases and DNase. In this situation, the fungus to be evolved can acquire DNA by exchange or transformation from another strain known to have the desired property. Alternatively, the fungus to be evolved can be subject to random mutagenesis. The fungus to be evolved is stochastic &/or non-stochastic mutagenized as above. The presence of such enzymes, or lack thereof, can be assayed by contacting the culture media from individual isolates with a fluorescent molecule tethered to a support via a peptide or DNA linkage. Cleavage of the linkage releases detectable fluorescence to the media.

4.9.12 ALTERED TRANSPORTERS TO USE SECONDARY COMPONENTS

Another desired property is producing fungi with altered transporters (e. g., MDR). Such altered transporters are useful, for example, in fungi that have been evolved to produce new secondary metabolites, to allow entry of precursors required for synthesis of the new secondary metabolites into a cell, or to allow efflux of the secondary metabolite from the cell. Transporters can be evolved by introduction of a library of transporter variants into fungal cells and allowing the cells to recombine by sexual or parasexual recombination. To evolve a transporter with capacity to transport a precursor into the cells, cells are propagated in the present of precursor, and cells are then screened for production of metabolite. To evolve a transporter with capacity to export a metabolite, cells are propagated under conditions supporting production of the metabolite, and screened for export of metabolite to culture medium.

A general method of fungal stochastic &/or non-stochastic mutagenesis is shown herein. Spores from a frozen stock, a lyophilized stock, or fresh from an agar plate are used to inoculate suitable liquid medium (1). Spores are germinated resulting in hyphal growth (2). Mycelia are harvested, and washed by filtration and/or centrifugation. Optionally the sample is pretreated with DTT to enhance protoplast formation (3). Protoplasting is performed in an osmotically stablizing medium (e.g., 1 m NaCl/2OmM MgSO₄, pH 5.8) by the addition of cell wall- degrading enzyme (e.g., Novozyme 234) (4). Cell wall degrading enzyme is removed by repeated washing

with osmotically stabilizing solution (5). Protoplasts can be separated from mycelia, debris and spores by filtration through miracloth, and density centrifugation (6). Protoplasts are harvested by centrifugation and resuspended to the appropriate concentration. This step may lead to some protoplast fusion (7). Fusion can be stimulated by addition of PEG (e.g., PEG 3350), and/or repeated centrifugation and resuspension with or without PEG. Electrofusion can also be performed (8). Fused protoplasts can optionally be enriched from unfused protoplasts by sucrose gradient sedimentation (or other methods of screening described above). Fused protoplasts can optionally be treated with ultraviolet irradiation to stimulate recombination (9). Protoplasts are cultured on osmotically stabilized agar plates to regenerate cell walls and form mycelia (10). The mycelia are used to generate spores (11), which are used as the starting material in the next round of stochastic &/or non-stochastic mutagenesis (12). Selection for a desired property can be performed either on regenerated mycelia or spores derived therefrom.

In an alternative method, protoplasts are formed by inhibition of one or more enzymes required for cell wall synthesis. The inhibitor should be fungistatic rather than fungicidal under the conditions of use. Examples of inhibitors include antifungal compounds described by (e.g., Georgopapadakou & Walsh, *Antimicrob. Ag. Chemother.* 40, 279-291 (1996); Lyman & Walsh, *Drugs* 44, 9-35 (1992)). Other examples include chitin synthase inhibitors (polyoxin or nikkomycin compounds) and/or glucan synthase inhibitors (e.g. echinocandins, papulocandins, pneumocandins). Inhibitors should be applied in osmotically stabilized medium. Cells stripped of their cell walls can be fused or otherwise employed as donors or hosts in genetic transformation/strain development programs.

In a further variation, protoplasts are prepared using strains of fungi, which are genetically deficient or compromised in their ability to synthesize intact cell walls. Such mutants are generally referred to as fragile, osmotic-remedial, or cell wall-less, and are obtainable from strain depositories. Examples of such strains include

Neurospora crassa os mutants (Selitrennikoff, Antimicrob. Agents. Chemother. 23, 757-765 (1983)). Some such mutations are temperature-sensitive. Temperature-sensitive strains can be propagated at the permissive temperature for purposes of selection and amplification and at a nonpermissive temperature for purposes of protoplast formation and fusion. A temperature sensitive strain *Neurospora crassa* os strain has been described which propagates as protoplasts when growth in osmotically stabilizing medium containing sorbose and polyoxin at nonpermissive temperature but generates whole cells on transfer to medium containing sorbitol at a permissive temperature. See US 4,873,196.

Other suitable strains can be produced by targeted mutagenesis of genes involved in chitin synthesis, glucan synthesis and other cell wall-related processes. Examples of such genes include CHT1, CHT2 and CALI (or CSD2) of *Saccharomyces cerevisiae* and *Candida* spp. (Georgopapadakou & Walsh 1996); ETGI/FKSI/CNDI/CWH53/PB RI and homologs in *S. cerevisiae*, *Candida albicans*, *Cryptococcus neoformans*, *Aspergillus fumigatus*, ChvA1NdvA Agrobacterium and Rhizobium. Other examples are M4, orlB, orlC, MD, tsE, and bimG of *Aspergillus nidulans* (Borgia, J Bacteriol. 174, 3 77-3 89 (1992)).

Strains of *A. nidulans* containing OrlA1 or tsE1 mutations lyse at restrictive temperatures. Lysis of these strains may be prevented by osmotic stabilization, and the mutations may be complemented by the addition of N-acetylglucosamine (GlcNac). BimG11 mutations are ts for a type 1 protein phosphatase (germlines of strains carrying this mutation lack chitin, and conidia swell and lyse). Other suitable genes are chsA, chsB, chsC, chsD and chsE of *Aspergillus fumigatus*; chs1 and chs2 of *Neurospora crassa*; *Phycomyces blakesleeanus* MM and chs 1, 2 and 3 of *S. cerevisiae*. Chs 1 is a non-essential repair enzyme; chs2 is involved in septum formation and chs3 is involved in cell wall maturation and bud ring formation.

Other useful strains include *S. cerevisiae* CLY (cell lysis) mutants such as ts strains (Paravicini et al., Mol. Cell Biol 12, 4896-4905 (1992)), and the CLY 15 strain which harbors a PKC 1 gene deletion. Other useful strains include strain VY 1160 containing a ts mutation in *srb* (encoding actin) (Schade et al. Acta Histochem. Suppl. 41, 193-200 (1991)), and a strain with an *ses* mutation which results in increased sensitivity to cell-wall digesting enzymes isolated from snail gut (Metha & Gregory, AppL Environ. Microbiol 41, 992-999 (1981)). Useful strains of *C albicans* include those with mutations in *chsl*, *chs2*, or *chs3* (encoding chitin synthetases), such as osmotic remedial conditional lethal mutants described by Payton & de Tiani, Curr. Genet. 17, 293-296 (1990); *C. utilis* mutants with increased sensitivity to cell-wall digesting enzymes isolated from snail gut (Metha & Gregory, 1981, supra); and *X crassa* mutants os-1, os-2, os-3, os-4, os-5, and os-6. See, Selitrennikoff, Antimicrob. Agents Chemother. 23, 757-765 (1983). Such mutants grow and divide without a cell wall at 37 C, but at 22 C produce a cell wall.

Targeted mutagenesis can be achieved by transforming cells with a positive-negative selection vector containing homologous regions flanking a segment to be targeted, a positive selection marker between the homologous regions and a negative selection marker outside the homologous regions (see Capecchi, US 5,627,059). In a variation, the negative selection marker can be an antisense transcript of the positive selection marker (see US 5,527,674).

Other suitable cells can be selected by random mutagenesis or stochastic &/or non-stochastic mutagenesis procedures in combination with selection. For example, a first subpopulation of cells are mutagenized, allowed to recover from mutagenesis, subjected to incomplete degradation of cell walls and then contacted with protoplasts of a second subpopulation of cells. Hybrids cells bearing markers from both subpopulations are identified (as described above) and used as the starting materials in a subsequent round of stochastic &/or non-stochastic mutagenesis. This selection scheme selects both for cells with capacity for spontaneous protoplast formation and

for cells with enhanced recombinogenicity.

In a further variation, cells having capacity for spontaneous protoplast formation can be crossed with cells having enhanced recombinogenicity evolved using other methods of the invention. The hybrid cells are particularly suitable hosts for whole genome stochastic &/or non-stochastic mutagenesis.

Cells with mutations in enzymes involved in cell wall synthesis or maintenance can undergo fusion simply as a result of propagating the cells in osmotic-protected culture due to spontaneous protoplast formation. If the mutation is conditional, cells are shifted to a nonpermissive condition. Protoplast formation and fusion can be accelerated by addition of promoting agents, such as PEG or an electric field (See Philipova & Venkov, Yeast 6, 205-212 (1990); Tsoneva et al., FFMS Microbiol Lett. 51, 61-65 (1989)).

4.10 PROCESS OF SEXUAL REPRODUCTION

Sexual reproduction provides a mechanism for stochastic &/or non-stochastic mutagenesis genetic material between cells. A sexual reproductive cycle is characterized by an alteration of a haploid phase and a diploid phase. Diploidy occurs when two haploid gamete nuclei fuse (karyogamy). The gamete nuclei can come from the same parental strains (self-fertile), such as in the homothallic fungi. In heterothallic fungi, the parental strains come from strains of different mating type.

A diploid cell converts to haploidy via meiosis, which essentially consists of two divisions of the nucleus accompanied by one division of the chromosomes. The products of one meiosis are a tetrad (4 haploid nuclei). In some cases, a mitotic division occurs after meiosis, giving rise to eight product cells. The arrangement of the resultant cells (usually enclosed in spores) resembles that of the parental strains. The length of the haploid and diploid stages differs in various fungi: for example, the Basidiomycetes and many of the Ascomycetes have a mostly haploid life cycle (that

is, meiosis occurs immediately after karyogamy), whereas others (e.g., *Saccharomyces cerevisiae*) are diploid for most of their life cycle (karyogamy occurs soon after meiosis). Sexual reproduction can occur between cells in the same strain (selfing) or between cells from different strains (outcrossing).

Sexual dimorphism (dioecism) is the separate production of male and female organs on different mycelia. This is a rare phenomenon among the fungi, although a few examples are known. Heterothallism (one locus-two alleles) allows for outcrossing between crosscompatable strains which are self-incompatable. The simplest form is the two allele-one locus system of mating types/factors, illustrated by the following organisms: A and a in *Neurospora*; a and - in *Saccharomyces*, plus and minus in *Schizzosaccharomyces* and *Zygomycetes*; 1 and 2 in *Ustilago*.

Multiple-allelomorph heterothallism is exhibited by some of the higher Basidiomycetes (e.g. Gasteromycetes and Hymenomycetes), which are heterothallic and have several mating types determined by multiple alleles. Heterothallism. In these organisms is either bipolar with one mating type factor, or tetrapolar with two unlinked factors, A and B. Stable, fertile heterokaryon formation depends on the presence of different A factors and, in the case of tetrapolar organisms, of different B factors as well. This system is effective in the promotion of outbreeding and the prevention of self-breeding. The number of different mating factors may be very large (i.e. thousands) (Kothe, FEMS Microbiol Rev. 18, 65-87 (1996)), and non-parental mating factors may arise by recombination.

4.10.1 INTRODUCING SEXUAL CYCLES

4.10.2 MEIOSIS

4.10.2.1 HETEROKARYON-A CELL OR HYPHA CONTAINING TWO OR MORE NUCLEI OF DIFFERENT GENETIC CONSTITUTIONS

One desired property is the introduction of meiotic apparatus into fungi presently lacking a sexual cycle (see Sharon et al., Mol. Gen. Genet. 251, 60-68 (1996)). A scheme for introducing a sexual cycle into the fungi *P. chrysogenum* (a fungus imperfecti) is shown herein. Subpopulations of protoplasts are formed from *A. nidulans* (which has a sexual cycle) and *P. chrysogenum*, which does not. The two strains preferably bear different markers. The *A. nidulans* protoplasts are killed by treatment with UV or hydroxylamine. The two subpopulations are fused to form heterokaryons. In some heterokaryons, nuclei fuse, and some recombination occurs. Fused cells are cultured under conditions to generate new cell walls and then to allow sexual recombination to occur. Cells with recombinant genomes are then selected (e.g., by selecting for complementation of auxotrophic markers present on the respective parent strains). Cells with hybrid genomes are more likely to have acquired the genes necessary for a sexual cycle. Protoplasts of cells can then be crossed with killed protoplasts of a further population of cells known to have a sexual cycle (the same or different as the previous round) in the same manner, followed by selection for cells with hybrid genomes.

4.10.2.2 VEGETATIVE COMPATIBILITY BETWEEN CLASSES OF FUNGI

Within the above four classes, fungi are also classified by vegetative compatibility group. Fungi within a vegetative compatibility group can form heterokaryons with each other. Thus, for exchange of genetic material between different strains of fungi, the fungi are usually prepared from the same vegetative compatibility group. However, some genetic exchange can occur between fungi from different incompatibility groups as a result of parasexual reproduction (see Timberlake et al., US 5,605,820). Further, as discussed elsewhere, the natural vegetative compatibility group of fungi can be expanded as a result of stochastic &/or non-stochastic mutagenesis.

Several isolates of *Aspergillus nidulans*, *A. flavus*, *A. fumigatus*, *Penicillium chrysogenum*, *P. notatum*, *Cephalosporium chrysogenum*, *Neurospora crassa*,

Aureobasidium pullulans have been karyotyped. Genome sizes generally range between 20 and 50 Mb among the Aspergilli. Differences in karyotypes often exist between similar strains and are also caused by transformation with exogenous DNA. Filamentous fungal genes contain introns, usually ~50-100 bp in size, with similar consensus 5' and 3' splice sequences. Promotion and termination signals are often cross-recognizable, enabling the expression of a gene/pathway from one fungus (e.g. *A. nidulans*) in another (e.g. *P. chrysogenum*). The major components of the fungal cell wall are chitin (or chitosan), beta- glucan, and mannoproteins. Chitin and beta-glucan form the scaffolding, mannoproteins are interstitial components which dictate the wall's porosity, antigenicity and adhesion. Chitin synthetase catalyzes the polymerization of beta-(1,4)-linked N-acetylglucosamine (GlcNAc) residues, forming linear strands running antiparallel; beta-(1,3)-glucan synthetase catalyze the homopolymerization of glucose.

4.11 EVOLUTION

4.11.1 ARTIFICIALLY EVOLVING CELLS TO ACQUIRE A NEW OR IMPROVED PROPERTY BY STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS

The invention provides a number of strategies for evolving metabolic and bioprocessing pathways through the technique of recursive sequence recombination. One strategy entails evolving genes that confer the ability to use a particular substrate of interest as a nutrient source in one species to confer either more efficient use of that substrate in that species, or comparable or more efficient use of that substrate in a second species. Another strategy entails evolving genes that confer the ability to detoxify a compound of interest in one or more species of organisms. Another strategy entails evolving new metabolic pathways by evolving an enzyme or metabolic pathway for biosynthesis or degradation of a compound A related to a compound B for the ability to biosynthesize or degrade compound B, either in the host of origin or a new host. A further strategy entails evolving a gene or metabolic pathway for more efficient or optimized expression of a particular metabolite or gene

product. A further strategy entails evolving a host/vector system for expression of a desired heterologous product. These strategies may involve using all the genes in a multi-step pathway, one or several genes, genes from different organisms, or one or more fragments of a gene.

The strategies generally entail evolution of gene(s) or segment(s) thereof to allow retention of function in a heterologous cell or improvement of function in a homologous or heterologous cell. Evolution is effected generally by a process termed recursive sequence recombination. Recursive sequence recombination can be achieved in many different formats and permutations of formats, as described in further detail below. These formats share some common principles. Recursive sequence recombination entails successive cycles of recombination to generate molecular diversity, i.e., the creation of a family of nucleic acid molecules showing substantial sequence identity to each other but differing in the presence of mutations. Each recombination cycle is followed by at least one cycle of screening or selection for molecules having a desired characteristic. The molecule(s) selected in one round form the starting materials for generating diversity in the next round. In any given cycle, recombination can occur *in vivo* or *in vitro*. Furthermore, diversity resulting from recombination can be augmented in any cycle by applying prior methods of mutagenesis (e.g., error-prone PCR or cassette mutagenesis, passage through bacterial mutator strains, treatment with chemical mutagens) to either the substrates for or products of recombination.

4.11.2 BASIC APPROACH

4.11.2.1 SUCCESSIVE CYCLES OF RECOMBINATION AND SCREENING/SELECTION

The invention provides methods for artificially evolving cells to acquire a new or improved property by recursive sequence recombination. Briefly, recursive sequence recombination entails successive cycles of recombination to generate molecular diversity and screening/selection to take advantage of that molecular diversity. That is, a family of nucleic acid molecules is created showing substantial sequence and/or

structural identity but differing as to the presence of mutations. These sequences are then recombined in any of the described formats so as to optimize the diversity of mutant combinations represented in the resulting recombined library. Typically, any resulting recombinant nucleic acids or genomes are recursively recombined for one or more cycles of recombination to increase the diversity of resulting products. After this recursive recombination procedure, the final resulting products are screened and/or selected for a desired trait or property.

Alternatively, each recombination cycle can be followed by at least one cycle of screening or selection for molecules having a desired characteristic. In this embodiment, the molecule(s) selected in one round form the starting materials for generating diversity in the next round.

The cells to be evolved can be bacteria, archaebacteria, or eukaryotic cells and can constitute a homogeneous cell line or mixed culture. Suitable cells for evolution include the bacterial and eukaryotic, cell lines commonly used in genetic engineering, protein expression, or the industrial production or conversion of proteins, enzymes, primary metabolites, secondary metabolites, fine, specialty or commodity chemicals. Suitable mammalian cells include those from, e.g., mouse, rat, hamster, primate, and human, both cell lines and primary cultures. Such cells include stem cells, including embryonic stem cells and hemopoietic stem cells, zygotes, fibroblasts, lymphocytes, Chinese hamster ovary (CHO), mouse fibroblasts (NIHM), kidney, liver, muscle, and skin cells. Other eukaryotic cells of interest include plant cells, such as maize, rice, wheat, cotton, soybean, sugarcane, tobacco, and arabidopsis; fish, algae, fungi (penicillium, aspergillus, podospora, neurospora, saccharomyces), insect (e.g., baculo lepidoptera), yeast (picchia and saccharomyces, Schizosaccharomycespombe). Also of interest are many bacterial cell types, both gram- negative and gram-positive, such as *Bacillus subtilis*, *B. licehniformis*, *B. cereus*, *Escherichia coli*, *Streptomyces*, *Pseudomonas*, *Salmonella*, *Actinomycetes*, *Lactobacillus*, *Acelonitcbacter*, *Deinococcus*, and *Erwinia*. The complete genome sequences of *E. coli* and *Bacillus subtilis* are described by Blattner et al., *Science* 277, 1454- 1462 (1997); Kunst et al., *Nature* 390, 249-256 (1997).

4.11.2.1.1 GOAL IS TO ACHIEVE VARIATION

Evolution commences by generating a population of variant cells. Typically, the cells in the population are of the same type but represent variants of a progenitor cell. In some instances, the variation is natural as when different cells are obtained from different individuals within a species, from different species or from different genera. In other instances, variation is induced by mutagenesis of a progenitor cell.

Mutagenesis can be effected by subjecting the cell to mutagenic agents, or if the cell is a mutator cell (e.g., has mutations in genes involved in DNA replication, recombination and/or repair which favor introduction of mutations) simply by propagating the mutator cells. Mutator cells can be generated from successive selections for simple phenotypic changes (e.g., acquisition of rifampicin-resistance, then nalidixic acid resistance then lac- to lac+ (see Mao et al., J Bacteriol 179, 417-422 (1997)), or mutator cells can be generated by exposure to specific inhibitors of cellular factors that result in the mutator phenotype. These could be inhibitors of mutS, mutL, mutD, recD, mutY, mutM, dam, uvrD and the like.

More generally, mutations are induced in cell populations using any available mutation technique. Common mechanisms for inducing mutations include, but are not limited to, the use of strains comprising mutations such as those involved in mismatch repair. e.g. mutations in mutS, mutT, mutL and mutH; exposure to UV light; Chemical mutagenesis, e.g. use of inhibitors of MMR, DNA damage inducible genes, or SOS inducers; overproduction/ underproduction/ mutation of any component of the homologous recombination complex/pathway, e.g. RecA, ssb, etc. overproduction/ underproduction/ mutation of genes involved in DNA synthesis/homeostasis; overproduction/ underproduction/ mutation of recombination-stimulating genes from bacteria, phage (e.g. Lambda Red function), or other organisms; addition of chi sites into/flanking the donor DNA fragments; coating the DNA fragments with RecA/ssb and the like.

In other instances, variation is the result of transferring a library of DNA fragments into the cells (e.g., by conjugation, protoplast fusion, liposome fusion, transformation, transduction or natural competence). At least one, and usually many of the fragments in the library, show some, but not complete, sequence or structural identity with a cognate or allelic gene within the cells sufficient to allow homologous recombination to occur.

For example, in one embodiment, homologous integration of a plasmid carrying a stochastic &/or non-stochastic mutagenized gene or metabolic pathway leads to insertion of the plasmid-borne sequences adjacent to the genomic copy. Optionally, a counter-selectable marker strategy is used to select for recombinants in which recombination occurred between the homologous sequences, leading to elimination of the counter-selectable marker. A variety of selectable and counter selectable markers are amply illustrated in the art. For a fist of useful markers, see, Berg and Berg (1996), Transposable element tools for microbial genetics, *Escherichia coli* and *Salmonella* Neidhardt. Washington, D.C., ASM Press. 2: 2588-2612; La Rossa, *ibid.*, 2527-2587. This strategy can be recursively repeated to maximize sequence diversity of targeted genes prior to screening/ selection for a desired trait or property.

The library of fragments can derive from one or more sources. One source of fragments is a genomic library of fragments from a different species, cell type, organism or individual from the cells being transfected. In this situation, many of the fragments in the library have a cognate or allelic gene in the cells being transformed but differ from that gene due to the presence of naturally occurring species variation, polymorphisms, mutations, and the presence of multiple copies of some homologous genes in the genome. Alternatively, the library can be derived from DNA from the same cell type as is being transformed after that DNA has been subject to induced mutation, by conventional methods, such as radiation, error-prone PCR, growth in a mutator organism, transposon mutagenesis, or cassette mutagenesis.

Alternatively, the library can derive from a genomic library of fragments generated from the pooled genomic DNA of a population of cells having the desired characteristics. Alternatively, the library can derive from a genomic library of fragments generated from the pooled genomic DNA of a population of cells having desired characteristics.

In any of these situations, the genomic library can be a complete genomic library or subgenome & library deriving, for example, from a selected chromosome, or part of a chromosome or an episomal element within a cell. As well as, or instead of these sources of DNA fragments, the library can contain fragments representing natural or selected variants of selected genes of known function (i.e., focused libraries).

The number of fragments in a library can vary from a single fragment to about 10^{10} with libraries having from 10^3 to 10^8 fragments being common. The fragments should be sufficiently long that they can undergo homologous recombination and sufficiently short that they can be introduced into a cell, and if necessary, manipulated before introduction. Fragment sizes can range from about 10 b to about 20mb. Fragments can be double- or single-stranded. The fragments can be introduced into cells as whole genomes or as components of viruses, plasmids, YACs, HACs or BACs or can be introduced as they are, in which case all or most of the fragments lack an origin of replication. Use of viral fragments with single-stranded genomes offer the advantage of delivering fragments in single stranded form, which promotes recombination. The fragments can also be joined to a selective marker before introduction. Inclusion of fragments in a vector having an origin of replication affords a longer period of time after introduction into the cell in which fragments can undergo recombination with a cognate gene before being degraded or selected against and lost from the cell, thereby increasing the proportion of cells with recombinant genomes. Optionally, the vector is a suicide vector capable of a longer existence than an isolated DNA fragment but not capable of permanent retention in the cell line. Such a vector can transiently express a marker for a sufficient time to screen for or select a cell bearing the vector (e.g., because cells transduced by the vector are the target cell type to be screened in subsequent selection assays), but is then degraded or otherwise rendered incapable of

expressing the marker. The use of such vectors can be advantageous in performing optional subsequent rounds of recombination to be discussed below. For example, some suicide vectors express a long- lived toxin which is neutralized by a short-lived molecule expressed from the same vector. Expression of the toxin alone will not allow vector to be established. Jense & Gerdes, Mol. Microbiol, 17, 205-210 (1995); Bernard et al., Gene 162, 159-160. Alternatively, a vector can be rendered suicidal by incorporation of a defective origin of replication (e.g. a temperature- sensitive origin of replication) or by omission of an origin of replication. Vectors can also be rendered suicidal by inclusion of negative selection markers, such as *ura3* in yeast or *sacB* in many bacteria.

These genes become toxic only in the presence of specific compounds. Such vectors can be selected to have a wide range of stabilities. A list of conditional replication defects for vectors which can be used, e.g., to render the vector replication defective is found, e.g., in Berg and Berg (1996), "Transposable element tools for microbial genetics" Escherichia coli and Salmonella Neidhardt. Washington, D.C., ASM Press. 2: 2588-2612. Similarly, a fist of counter selectable markers, generally applicable to vector selection is also found in Berg and Berg, id. See also, LaRossa (1996), "Mutant selections linking physiology, inhibitors, and genotypes" Escherichia coli and Salmonella F. C. Neidhardt. Washington, D.C., ASM Press. 2: 2527-2587.

After introduction into cells, the fragments can recombine with DNA present in the genome, or episomes of the cells by homologous, nonhomologous or site-specific recombination. For present purposes, homologous recombination makes the most significant contribution to evolution of the cells because this form of recombination amplifies the existing diversity between the DNA of the cells being transfected and the DNA fragments. For example, if a DNA fragment being transfected differs from a cognate or allelic gene at two positions, there are four possible recombination products, and each of these recombination products can be formed in different cells in the transformed population. Thus, homologous recombination of the fragment doubles

the initial diversity in this gene. When many fragments recombine with corresponding cognate or allelic genes, the diversity of recombination products with respect to starting products increases exponentially with the number of mutations.

Recombination results in modified cells having modified genomes and/or episomes. Recursive recombination prior to selection further increases diversity of resulting modified cells.

The variant cells, whether the result of natural variation, mutagenesis, or recombination are screened or selected to identify a subset of cells that have evolved toward acquisition of a new or improved property. The nature of the screen, of course, depends on the property and several examples will be discussed below. Typically, recombination is repeated before initial screening. Optionally, however, the screening can also be repeated before performing subsequent cycles of recombination.

Stringency can be increased in repeated cycles of screening. The subpopulation of cells surviving screening are optionally subjected to a further round of recombination. In some instances, the further round of recombination is effected by propagating the cells under conditions allowing exchange of DNA between cells. For example, protoplasts can be formed from the cells, allowed to fuse, and regenerated. Cells with recombinant genomes are propagated from the fused protoplasts. Alternatively, exchange of DNA can be promoted by propagation of cells or protoplasts in an electric field. For cells having a conjugative transfer apparatus, exchange of DNA can be promoted simply by propagating the cells.

4.11.2.1.2 USING TWO SEPARATE POOLS: AMPLIFY FIRST POOL AND ADD TO SECOND POOL

In other methods, the further round of recombination is performed by a split and pool approach. That is, the surviving cells are divided into two pools. DNA is isolated from one pool, and if necessary amplified, and then transformed into the other pool.

Accordingly, DNA fragments from the first pool constitute a further library of

fragments and recombine with cognate fragments in the second pool resulting in further diversity. As shown, a pool of mutant bacteria with improvements in a desired phenotype is obtained and split. Genes are obtained from one half, e.g., by PCR, by cloning of random genomic fragments, by infection with a transducing phage and harvesting transducing particles, or by the introduction of an origin of transfer (OriT) randomly into the relevant chromosome to create a donor population of cells capable of transferring random fragments by conjugation to an acceptor population. These genes are then stochastic &/or non-stochastic mutagenized (in vitro by known methods or in vivo as taught herein), or simply cloned into an allele replacement vector (e.g., one carrying selectable and counter-selectable markers). The gene pool is then transformed into the other half of the original mutant pool and recombinants are selected and screened for further improvements in phenotype. These best variants are used as the starting point for the next cycle. Alternatively, recursive recombination by any of the methods noted can be performed prior to screening, thereby increasing the diversity of the population of cells to be screened.

4.11.2.1.3 SURVIVING CELLS ARE TRANSFECTED INTO FRESH DNA

In other methods, some or all of the cells surviving screening are transfected with a fresh library of DNA fragments, which can be the same or different from the library used in the first round of recombination. In this situation, the genes in the fresh library undergo recombination with cognate genes in the surviving cells. If genes are introduced as components of a vector, compatibility of this vector with any vector used in a previous round of transfection should be considered. If the vector used in a previous round was a suicide vector, there is no problem of incompatibility. If, however, the vector used in a previous round was not a suicide vector, a vector having a different incompatibility origin should be used in the subsequent round. In all of these formats, further recombination generates additional diversity in the DNA component of the cells resulting in further modified cells.

The further modified cells are subjected to another round of screening/selection according to the same principles as the first round. Screening/selection identifies a

subpopulation of further modified cells that have further evolved toward acquisition of the property. This subpopulation of cells can be subjected to further rounds of recombination and screening according to the same principles, optionally with the stringency of screening being increased at each round. Eventually, cells are identified that have acquired the desired property.

4.11.3 VARIATIONS

4.11.3.1 COATING WITH RecA TO ENRICH DIVERSITY OF HOMOLOGOUS RECOMBINATION

The frequency of homologous recombination between library fragments and cognate endogenous genes can be increased by coating the fragments with a recombinogenic protein before introduction into cells. See Pati et al., Molecular Biology of Cancer 1, 1 (1996); Sena & Zarling, Nature Genetics 3, 365 (1996); Revet et al., J Mol. Biol. 232, 779- 791 (1993); Kowalczykowski & Zarling in Gene Targeting (CRC 1995), Ch. 7. The recombinogenic protein promotes homologous pairing and/or strand exchange. The best characterized recA protein is from *E. coli* and is available from Pharmacia (Piscataway, NJ).

In addition to the wild-type protein, a number of mutant recA-like proteins have been identified (e.g., recA803). Further, many organisms have recA-like recombinases with strand- transfer activities (e.g., Ogawa et al., Cold Spring Harbor Symposium on Quantitative Biology 18, 567-576 (1993); Johnson & Symington, Mol. Cell. Biol. 15, 4843-4850 (1995); Fugisawa et al., Nucl. Acids Res. 13, 7473 (1985); Hsieh et al., Cell 44, 885 (1986); Hsieh et al., J Biol. Chem. 264, 5089 (1989); Fishel et al., Proc. Nail. Acad. Sci. USA 85, 3683 (1988); Cassuto et al., Mol. Gen. Genet. 208, 10 (1987); Ganea et al., Mol. Cell Biol. 7, 3124 (1987); Moore et al., J Biol. Chem. 19, 11108 (1990); Keene et al., Nucl. Acids Res. 12, 3057 (1984); Kimiec, Cold Spring Harbor Symp. 48, 675 (1984); Kimiec, Cell 44, 545 (1986); Kolodner et al., Proc. Nad. Acad Sci. USA 84, 5560 (1987); Sugino et al., Proc. Natl Acad Sci. USA 85,

3683 (1985). Halbrook et al., J. Biol Chem. 264, 21403 (1989); Eisen et al., Proc. Natl. Acad. Sci. USA 85, 7481 (1988); McCarthy et al. , Proc. Natl. Acad Sci. USA 85, 5854 (1988). Lowenhaupt et al., J Biol. Chem. 264, 20568 (1989). Examples of such recombinase proteins include recA, recA803, uvsX, (Roca, A.I., Crit. Rev. Biochem. Molec. Biol. 25, 415 (1990)), sepI (Kolodner et al., Proc. Natl. Acad. Sci. (U.S.A.) 84, 5560 (1987); Tishkoff et al., Molec. Cell. Biol 11, 2593), RuvC (Dunderdale et al., Nature 354, 506 (1991)), DS72, KEMI, XRATI (Dykstra et al., Molec. Cell. Biol 11, 25 83 (1991)), STP /DST1 (Clark et al., Molec. Cell. Biol 11, 2576 (1991)), HPP-I (Moore et al., Proc. Natl. Acad. Sci. (U.S.A.) 88, 9067 (1991)), other eukaryotic recombinases (Bishop et al., Cell 69, 439 (1992); Shinohara et al., Cell 69, 457. RecA protein forms a nucleoprotein filament when it coats a single-stranded DNA. In this nucleoprotein filament, one monomer of recA protein is bound to about 3 nucleotides. This property of recA to coat single-stranded DNA is essentially sequence independent, although particular sequences favor initial loading of recA onto a polynucleotide (e.g., nucleation sequences). The nucleoprotein filament(s) can be formed on essentially any DNA to be stochastic &/or non-stochastic mutagenized and can form complexes with both single-stranded and double-stranded DNA in prokaryotic and eukaryotic cells.

Before contacting with recA or other recombinase, fragments are often denatured, e.g., by heat-treatment. RecA protein is then added at a concentration of about 1-10 gM. After incubation, the recA-coated single-stranded DNA is introduced into recipient cells by conventional methods, such as chemical transformation or electroporation. In general, it can be desirable to coat the DNA with a RecA homolog isolated from the organism into which the coated DNA is being delivered. Recombination involves several cellular factors and the host RecA equivalent generally interacts better with other host factors than less closely related RecA molecules. The fragments undergo homologous recombination with cognate endogenous genes. Because of the increased frequency of recombination due to recombinase coating, the fragments need not be introduced as components of vectors. Fragments are sometimes coated with other nucleic acid binding proteins that promote

recombination, protect nucleic acids from degradation, or target nucleic acids to the nucleus. Examples of such proteins includes Agrobacterium virE2 (Duffenberger et al., Proc. Natl. Acad. Sci. USA 86, 9154-9158 (1989)). Alternatively, the recipient strains are deficient in RecD activity. Single stranded ends can also be generated by 3'-5' exonuclease activity or restriction enzymes producing 5' overhangs.

4.11.3.2 AFFINITY CHROMATOGRAPHY WITH MutS TO ENRICH FOR FRAGMENTS HAVING AT LEAST ONE MISMATCH

The *E. coli* mismatch repair protein MutS can be used in affinity chromatography to enrich for fragments of double-stranded DNA containing at least one base of mismatch. The MutS protein recognizes the bubble formed by the individual strands about the point of the mismatch. See, e.g., Hsu & Chang, WO 9320233. The strategy of affinity enriching for partially mismatched duplexes can be incorporated into the present methods to increase the diversity between an incoming library of fragments and corresponding cognate or allelic genes in recipient cells.

MutS is used to increase diversity. The DNA substrates for enrichment are substantially similar to each other but differ at a few sites.

For example, the DNA substrates can represent complete or partial genomes (e.g., a chromosome library) from different individuals with the differences being due to polymorphisms. The substrates can also represent induced mutants of a wild type sequence.

The DNA substrates are pooled, restriction digested, and denatured to produce fragments of single-stranded DNA. The single-stranded DNA is then allowed to reanneal. Some single-stranded fragments reanneal with a perfectly matched complementary strand to generate perfectly matched duplexes. Other single-stranded fragments anneal to generate mismatched duplexes. The mismatched duplexes are enriched from perfectly matched duplexes by MutS chromatography (e.g., with MutS

immobilized to beads). The mismatched duplexes recovered by chromatography are introduced into recipient cells for recombination with cognate endogenous genes as described above. MutS affinity chromatography increases the proportion of fragments differing from each other and the cognate endogenous gene. Thus, recombination between the incoming fragments and endogenous genes results in greater diversity.

A second strategy for MutS enrichment. In this strategy, the substrates for MutS enrichment represent variants of a relatively short segment, for example, a gene or cluster of genes, in which most of the different variants differ at no more than a single nucleotide. The goal of MutS enrichment is to produce substrates for recombination that contain more variations than sequences occurring in nature. This is achieved by fragmenting the substrates at random to produce overlapping fragments. The fragments are denatured and reannealed as in the first strategy. Reannealing generates some mismatched duplexes which can be separated from perfectly matched duplexes by MutS affinity chromatography. As before, MutS chromatography enriches for duplexes bearing at least a single mismatch. The mismatched duplexes are then stochastic &/or non-stochastic mutagenized into longer fragments. This is accomplished by cycles of denaturation, reannealing, and chain extension of partially annealed duplexes. After several such cycles, fragments of the same length as the original substrates are achieved, except that these fragments differ from each other at multiple sites. These fragments are then introduced into cells where they undergo recombination with cognate endogenous genes.

4.11.3.3 SUICIDE VECTOR ENRICHES MUTATIONS FOR CELLS THAT HAVE INTEGRATED THE VECTOR INTO THE HOST CHROMOSOME

The invention further provides methods of enriching for cells bearing modified genes relative to the starting cells. This can be achieved by introducing a DNA fragment library (e.g., a single specific segment or a whole or partial genomic library) in a suicide vector (i.e., lacking a functional replication origin in the recipient cell type) containing both positive and negative selection markers. Optionally, multiple fragment libraries from different sources (e.g., *B. subtilis*, *B. lichenifonnis* and *B. cereus*) can be cloned into different vectors bearing different selection markers.

Suitable positive selection markers include neo^R, kanamycin^R, hyg, hisD, gpt, ble,

tet^R. Suitable negative selection markers include hsv-tk, hprt, gpt, SacB ura3 and cytosine deaminase. A variety of examples of conditional replication vectors, mutations affecting vector replication, limited host range vectors, and counterselectable markers are found in Berg and Berg, *supra*, and LaRossa, *ibid.* and the references therein.

In one example, a plasmid with R6K and f1 origins of replication, a positively selectable marker (beta-lactamase), and a counterselectable marker (*B. subtilis* sacB) was used. M 13 transduction of plasmids containing cloned genes were efficiently recombined into the chromosomal copy of that gene in a rep mutant *E. coli* strain.

Another strategy for applying negative selection is to include a wild type rpsL gene (encoding ribosomal protein S12) in a vector for use in cells having a mutant rpsL gene conferring streptomycin resistance. The mutant form of rpsL is recessive in cells having wild type rpsL. Thus, selection for Sm resistance selects against cells having a wild type copy of rpsL. See Skorupski & Taylor, *Gene* 169, 47-52 (1996).

Alternatively, vectors bearing only a positive selection marker can be used with one round of selection for cells expressing the marker, and a subsequent round of screening for cells that have lost the marker (e.g., screening for drug sensitivity). The screen for cells that have lost the positive selection marker is equivalent to screening against expression of a negative selection marker. For example, *Bacillus* can be transformed with a vector bearing a CAT gene and a sequence to be integrated. See Harwood & Cutting, *Molecular Biological Methods for Bacillus*, at pp. 31-33. Selection for chloramphenicol resistance isolates cells that have taken up vector. After a suitable period to allow recombination, selection for CAT sensitivity isolates cells which have lost the CAT gene. About 50% of such cells will have undergone recombination with the sequence to be integrated.

Suicide vectors bearing a positive selection marker and optionally, a negative

selection marker and a DNA fragment can integrate into host chromosomal DNA by a single crossover at a site in chromosomal DNA homologous to the fragment. Recombination generates an integrated vector flanked by direct repeats of the homologous sequence. In some cells, subsequent recombination between the repeats results in excision of the vector and either acquisition of a desired mutation from the vector by the genome or restoration of the genome to wild type.

In the present methods, after transfer of the gene library cloned in a suitable vector, positive selection is applied for expression of the positive selection marker. Because nonintegrated copies of the suicide vector are rapidly eliminated from cells, this selection enriches for cells that have integrated the vector into the host chromosome. The cells surviving positive selection can then be propagated and subjected to negative selection, or screened for loss of the positive selection marker. Negative selection selects against cells expressing the negative selection marker. Thus, cells that have retained the integrated vector express the negative marker and are selectively eliminated. The cells surviving both rounds of selection are those that initially integrated and then eliminated the vector. These cells are enriched for cells having genes modified by homologous recombination with the vector. This process diversifies by a single exchange of genetic information. However, if the process is repeated either with the same vectors or with a library of fragments generated by PCR of pooled DNA from the enriched recombinant population, resulting in the diversity of targeted genes being enhanced exponentially each round of recombination. This process can be repeated recursively, with selection being performed as desired.

4.11.3.4 EXPLOITING KNOWN INFORMATION SUCH AS MAP LOCATION OR FUNCTION

In general, the above methods do not require knowledge of the number of genes to be optimized, their map location or their function. However, in some instances, where this information is available for one or more gene, it can be exploited. For example, if the property to be acquired by evolution is enhanced recombination of cells, one gene

likely to be important is recA, even though many other genes, known and unknown, may make additional contributions. In this situation, the recA gene can be evolved, at least in part, separately from other candidate genes. The recA gene can be evolved by any of the methods of recursive recombination described in Section V. Briefly, this approach entails obtaining, diverse forms of a recA gene, allowing the forms to recombine, selecting recombinants having improved properties, and subjecting the recombinants to further cycles of recombination and selection. At any point in the individualized improvement of recA, the diverse forms of recA can be pooled with fragments encoding other genes in a library to be used in the general methods described herein. In this way, the library is seeded to contain a higher proportion of variants in a gene known to be important to the property sought to be acquired than would otherwise be the case.

In one example, a plasmid is constructed carrying a non-functional (mutated) version of a chromosomal gene such as URA3, where the wild-type gene confers sensitivity to a drug (in this case 5-fluoro orotic acid). The plasmid also carries a selectable marker (resistance to another drug such as kanamycin), and a library of recA variants. Transformation of the plasmid into the cell results in expression of the recA variants, some of which will catalyze homologous recombination at an increased rate. Those cells in which homologous recombination occurred are resistant to the selectable drug on the plasmid, and to 5-fluoro orotic acid because of the disruption of the chromosomal copy of this gene.

The recA variants which give the highest rates of homologous recombination are the most highly represented in a pool of homologous recombinants. The mutant recA genes can be isolated from this pool by PCR, re-stochastic &/or non-stochastic mutagenized, cloned back into the plasmid and the process repeated. Other sequences can be inserted in place of recA to evolve other components of the homologous recombination system.

4.11.3.5 USING OWN HARVEST OF CELLS SO NO IMPURITIES

In some stochastic &/or non-stochastic mutagenesis methods, DNA substrates are isolated from natural sources and are not easily manipulated by DNA modifying or polymerizing enzymes due to recalcitrant impurities, which poison enzymatic reactions. Such difficulties can be avoided by processing DNA substrates through a harvesting strain. The harvesting strain is typically a cell type with natural competence and a capacity for homologous recombination between sequences with substantial diversity (e.g., sequences exhibiting only 75% sequence identity). The harvesting strain bears a vector encoding a negative selection marker flanked by two segments respectively complementary to two segments flanking a gene or other region of interest in the DNA from a target organism. The harvesting strain is contacted with fragments of DNA from the target organism. Fragments are taken up by natural competence, or other methods described herein, and a fragment of interest from the target organism recombines with the vector of the harvesting strain causing loss of the negative selection marker. Selection against the negative marker allows isolation of cells that have taken up the fragment of interest.

Stochastic &/or non-stochastic mutagenesis can be carried out in the harvester strain (e.g., a RecE/T strain) or vector can be isolated from the harvester strain for in vitro stochastic &/or non-stochastic mutagenesis or transfer to a different cell type for in vivo stochastic &/or non-stochastic mutagenesis. Alternatively, the vector can be transferred to a different cell type by conjugation, protoplast fusion or electrofusion. An example of a suitable harvester strain is *Acinetobacter calcoaceticus* mutS. Melnikov and Youngman, (1999) Nucl Acid Res 27(4):1056-1062. This strain is naturally competent and takes up DNA in a nonsequence-specific manner. Also, because of the mutS mutation, this strain is capable of homologous recombination of sequences showing only 75% sequence identity.

4.12 FURTHER APPLICATIONS

4.12.1 IMPROVED RECOMBINANCY

One goal of whole cell evolution is to generate cells having improved capacity for recombination. Such cells are useful for a variety of purposes in molecular genetics including the *in vivo* formats of recursive sequence recombination described in Section V. Almost thirty genes (e.g., recA, recB, recC, recD, recE, recF, recG, recO, recQ, recR, recT, ruvA, ruvB, ruvC, sbcB, ssb, topA, gyrA and B, lig, polA, uvrD, E, recL, mutD, mutH, mutL, mutT, mutU, helD) and DNA sites (e.g., chi, recN, sbcC) involved in genetic recombination have been identified in *E. coli*, and cognate forms of several of these genes have been found in other organisms (e.g., rad51, rad55-rad57, Dmc1 in yeast (see Kowalczykowski et al., *Microbiol. Rev.* 58, 401-465 (1994); Kowalczykowski & Zarling, *supra*) and human homologs of Rad51 and Dmc1 have been identified (see Sandier et al., *Nucl. Acids Res.* 24, 2125-2132 (1996)). At least some of the *E. coli* genes, including recA are functional in mammalian cells, and can be targeted to the nucleus as a fusion with SV40 large T antigen nuclear targeting sequence (Reiss et al., *Proc. Natl. Acad. Sci. USA*, 93, 3094-3098 (1996)). Further, mutations in mismatch repair genes, such as mutL, mutS, mutH mutT relax homology requirements and allow recombination between more diverged sequences (Rayssiguier et al., *Nature* 342, 396-401 (1989)). The extent of recombination between divergent strains can be enhanced by impairing mismatch repair genes and stimulating SOS genes. Such can be achieved by use of appropriate mutant strains and/or growth under conditions of metabolic stress, which have been found to stimulate SOS and inhibit mismatch repair genes. Vulic et al., *Proc. Natl. Acad. Sci. USA* 94 (1997). In addition, this can be achieved by impairing the products of mismatch repair genes by exposure to selective inhibitors.

Starting substrates for recombination are selected according to the general principles described above. That is, the substrates can be whole genomes or fractions thereof containing recombination genes or sites. Large libraries of essentially random fragments can be seeded with collections of fragments constituting variants of one or more known recombination genes, such as recA. Alternatively, libraries can be formed by mixing variant forms of the various known recombination genes and sites.

4.12.2 EXPRESSION OF GFP INDICATES CELL IS CAPABLE OF HOMOLOGOUS RECOMBINATION

The library of fragments is introduced into the recipient cells to be improved and recombination occurs, generating modified cells. The recipient cells preferably contain a marker gene whose expression has been disabled in a manner that can be corrected by recombination. For example, the cells can contain two copies of a marker gene bearing mutations at different sites, which copies can recombine to generate the wild type gene. A suitable marker gene is green fluorescent protein. A vector can be constructed encoding one copy of GFP having stop codons near the N-terminus, and another copy of GFP having stop codons near the C-terminus of the protein. The distance between the stop codons at the respective ends of the molecule is 500 bp and about 25% of recombination events result in active GFP. Expression of GFP in a cell signals that a cell is capable of homologous recombination to recombine in between the stop codons to generate a contiguous coding sequence. By screening for cells expressing GFP, one enriches for cells having the highest capacity for recombination. The same type of screen can be used following subsequent rounds of recombination. However, unless the selection marker used in previous round(s) was present on a suicide vector, subsequent round(s) should employ a second disabled screening marker within a second vector bearing a different origin of replication or a different positive selection marker to vectors used in the previous rounds.

4.12.3 INCREASED GENOME COPY NUMBER SO MORE CHROMOSOMES PER BACTERIAL CELL TO MAKE EVOLUTION QUICKER

The majority of bacterial cells in stationary phase cultures grown in rich media contain two, four or eight genomes. In minimal medium the cells contain one or two genomes. The number of genomes per bacterial cell thus depends on the growth rate of the cell as it enters stationary phase. This is because rapidly growing cells contain multiple replication forks, resulting in several genomes in the cells after termination.

The number of genomes is strain dependent, although all strains tested have more than one chromosome in stationary phase. The number of genomes in stationary phase cells decreases with time. This appears to be due to fragmentation and degradation of entire chromosomes, similar to apoptosis in mammalian cells. This fragmentation of genomes in cells containing multiple genome copies results in massive recombination and mutagenesis. Useful mutants may find ways to use energy sources that will allow them to continue growing. Multigenome or gene-redundant cells are much more resistant to mutagenesis and can be improved for a selected trait faster.

Some cell types, such as *Deinococcus radians* (Daly and Minton J Bacteriol 177, 5495-5505 (1995)) exhibit polyploidy throughout the cell cycle. This cell type is highly radiation resistant due to the presence of many copies of the genome. High frequency recombination between the genomes allows rapid removal of mutations induced by a variety of DNA damaging agents.

A goal of the present methods is to evolve other cell types to have increased genome copy number akin to that of *Deinococcus radians*. Preferably, the increased copy number is maintained through all or most of its cell cycle in all or most growth conditions. The presence of multiple genome copies in such cells results in a higher frequency of homologous recombination in these cells, both between copies of a gene in different genomes within the cell, and between a genome within the cell and a transfected fragment. The increased frequency of recombination allows the cells to be evolved more quickly to acquire other useful characteristics.

Starting substrates for recombination can be a diverse library of genes only a few of which are relevant to genomic copy number, a focused library formed from variants of gene(s) known or suspected to have a role in genomic copy number or a combination of the two. As a general rule one would expect increased copy number would be achieved by evolution of genes involved in replication and cell septation

such that cell septation is inhibited without impairing replication. Genes involved in replication include *tus*, *xerC*, *xerD*, *dif*, *gyrA*, *gyrB*, *parE*, *parC*, *dif*, *TerA*, *TerB*, *TerC*, *TerD*, *TerE*, *TerF*, and genes influencing chromosome partitioning and gene copy number include *minD*, *mukA* (*tolC*), *mukB*, *mukC*, *mukD*, *spoOJ*, *spoIIIE* (Wake & Errington, *Annu. Rev. Genet.* 29, 41-67 (1995)). A useful source of substrates is the genome of a cell type such as *Deinococcus radians* known to have the desired phenotype of multigenomic copy number. As well as, or instead of, the above substrates, fragments encoding protein or antisense RNA inhibitors to genes known to be involved in cell septation can also be used. In nature, the existence of multiple genomic copies in a cell type would usually not be advantageous due to the greater nutritional requirements needed to maintain this copy number. However, artificial conditions can be devised to select for high copy number.

Modified cells having recombinant genomes are grown in rich media (in which conditions, multicopy number should not be a disadvantage) and exposed to a mutagen, such as ultraviolet or gamma irradiation or a chemical mutagen, e.g., mitomycin, nitrous acid, photoactivated psoralens, alone or in combination, which induces DNA breaks amenable to repair by recombination. These conditions select for cells having multicopy number due to the greater efficiency with which mutations can be excised. Modified cells surviving exposure to mutagen are enriched for cells with multiple genome copies. If desired, selected cells can be individually analyzed for genome copy number (e.g., by quantitative hybridization with appropriate controls). Some or all of the collection of cells surviving selection provide the substrates for the next round of recombination. In addition, individual cells can be sorted using a cell sorter for those cells containing more DNA, e.g., using DNA specific fluorescent compounds or sorting for increased size using light dispersion. Eventually cells are evolved that have at least 2, 4, 6, 8 or 10 copies of the genome throughout the cell cycle. In a similar manner, protoplasts can also be recombined.

4.12.4 EVOLVE SECRETION PATHWAYS FOR BETTER EFFICIENCY

4.12.5 EVOLVE TO MANUFACTURE DRUGS OR CHEMICALS

The protein (or metabolite) secretion pathways of bacterial and eukaryotic cells can be

evolved to export desired molecules more efficiently, such as for the manufacturing of protein pharmaceuticals, small molecule drugs or specialty chemicals. Improvements in efficiency are particularly desirable for proteins requiring multisubunit assembly (such as antibodies) or extensive posttranslational modification before secretion.

The efficiency of secretion may depend on a number of genetic sequences including a signal peptide coding sequence, sequences encoding protein(s) that cleave or otherwise recognize the coding sequence, and the coding sequence of the protein being secreted. The latter may affect folding of the protein and the ease with which it can integrate into and traverse membranes. The bacterial secretion pathway in *E. coli* include the SecA, SecB, SecE, SecD and SecF genes. In *Bacillus subtilis*, the major genes are secA, secD, secE, secF, secY, ffh, ftsY together with five signal peptidase genes (sipS, sipT, sipU, sipV and sipW) (Kunst et al, supra). For proteins requiring posttranslational modification, evolution of genes effecting such modification may contribute to improved secretion. Likewise genes with expression products having a role in assembly of multisubunit proteins (e.g., chaperonins) may also contribute to improved secretion.

Selection of substrates for recombination follows the general principles discussed above. In this case, the focused libraries referred to above comprise variants of the known secretion genes. For evolution of prokaryotic cells to express eukaryotic proteins, the initial substrates for recombination are often obtained at least in part from eukaryotic sources.

Incoming fragments can undergo recombination both with chromosomal DNA in recipient cells and with the screening marker construct present in such cells (see below). The latter form of recombination is important for evolution of the signal coding sequence incorporated in the screening marker construct. Improved secretion can be screened by the inclusion of marker construct in the cells being evolved. The marker construct encodes a marker gene, operably linked to expression sequences, and usually operably linked to a signal peptide coding sequence. The marker gene is

sometimes expressed as a fusion protein with a recombinant protein of interest. This approach is useful when one wants to evolve the recombinant protein coding sequence together with secretion genes.

4.12.6 EVOLVE SO PRODUCT IS TOXIC TO CELL UNLESS SECRETED

In one variation, the marker gene encodes a product that is toxic to the cell containing the construct unless the product is secreted. Suitable toxin proteins include diphtheria toxin and ricin toxin. Propagation of modified cells bearing such a construct selects for cells that have evolved to improve secretion of the toxin. Alternatively, the marker gene can encode a ligand to a known receptor, and cells bearing the ligand can be detected by FACS using labeled receptor. Optionally, such a ligand can be operably linked to a phospholipid anchoring sequence that binds the ligand to the cell membrane surface following secretion. In a further variation, secreted marker protein can be maintained in proximity with the cell secreting it by distributing individual cells into agar drops. This is done, e.g., by droplet formation of a cell suspension. Secreted protein is confined within the agar matrix and can be detected by e.g., FACS. In another variation, a protein of interest is expressed as a fusion protein together with beta-lactamase or alkaline phosphatase. These enzymes metabolize commercially available chromogenic substrates (e.g., X-gal), but do so only after secretion into the periplasm. Appearance of colored substrate in a colony of cells therefore indicates capacity to secrete the fusion protein and the intensity of color is related to the efficiency of secretion.

The cells identified by these screening and selection methods have the capacity to secrete increased amounts of protein. This capacity may be attributable to increased secretion and increased expression, or from increased secretion alone.

4.12.7 EVOLVE TO ACQUIRE INCREASED EXPRESSION OF RECOMBINANT PROTEIN

Expression Cells can also be evolved to acquire increased expression of a recombinant protein. The level of expression is, of course, highly dependent on the construct from which the recombinant protein is expressed and the regulatory sequences, such as the promoter, enhancer(s) and transcription termination site contained therein. Expression can also be affected by a large number of host genes having roles in transcription, posttranslational modification and translation. In addition, host genes involved in synthesis of ribonucleotide and amino acid monomers for transcription and translation may have indirect effects on efficiency of expression. Selection of substrates for recombination follows the general principles discussed above. In this case, focused libraries comprise variants of genes known to have roles in expression. For evolution of prokaryotic cells to express eukaryotic proteins, the initial substrates for recombination are often obtained, at least in part, from eukaryotic sources; that is eukaryotic genes encoding proteins such as chaperonins involved in secretion and/assembly of proteins. Incoming fragments can undergo recombination both with chromosomal DNA in recipient cells and with the screening marker construct present in such cells (see below).

Screening for improved expression can be effected by including a reporter construct in the cells being evolved. The reporter construct expresses (and usually secretes) a reporter protein, such as GFP, which is easily detected and nontoxic. The reporter protein can be expressed alone or together with a protein of interest as a fusion protein. If the reporter gene is secreted, the screening effectively selects for cells having either improved secretion or improved expression, or both.

4.12.8 EVOVLE PLANT CELLS TO ACQUIRE RESISTANCE

A further application of recursive sequence recombination is the evolution of plant cells, and transgenic plants derived from the same, to acquire resistance to pathogenic diseases (fungi, viruses and bacteria), insects, chemicals (such as salt, selenium, pollutants, pesticides, herbicides, or the like), including, e.g., atrazine or glyphosate, or to modify chemical composition, yield or the like. The substrates for recombination

can again be whole genomic libraries, fractions thereof or focused libraries containing variants of gene(s) known or suspected to confer resistance to one of the above agents. Frequently, library fragments are obtained from a different species to the plant being evolved.

The DNA fragments are introduced into plant tissues, cultured plant cells, plant microspores, or plant protoplasts by standard methods including electroporation (From et al., Proc. Natl. Acad. Sci. USA 82, 5824 (1985), infection by viral vectors such as cauliflower mosaic virus (CaMV) (Hohn et al., Molecular Biology of Plant Tumors, (Academic Press, New York, 1982) pp. 549-560; Howell, US 4,407,956), high velocity ballistic penetration by small particles with the nucleic acid either within the matrix of small beads or particles, or on the surface (Klein et al., Nature 327, 70-73 (1987)), use of pollen as vector (WO 85/01856), or use of Agrobacterium tumefaciens or A. rhizogenes carrying a T-DNA plasmid in which DNA fragments are cloned. The T-DNA plasmid is transmitted to plant cells upon infection by Agrobacterium tumefaciens, and a portion is stably integrated into the plant genome (Horsch et al., Science 233, 496-498 (1984); Fraley et al., Proc. Natl. Acad. Sci. USA 80, 4803 (1983)).

Diversity can also be generated by genetic exchange between plant protoplasts according to the same principles described below for fungal protoplasts. Procedures for formation and fusion of plant protoplasts are described by Takahashi et al., US 4,677,066; Akagi et al., US 5,360,725; Shimamoto et al., US 5,250,433; Cheney et al., US 5,426,040.

4.12.9 PLANT GENOME STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS

Plant genome stochastic &/or non-stochastic mutagenesis allows recursive cycles to be used for the introduction and recombination of genes or pathways that confer improved properties to desired plant species. Any plant species, including weeds and

wild cultivars, showing a desired trait, such as herbicide resistance, salt tolerance, pest resistance, or temperature tolerance, can be used as the source of DNA that is introduced into the crop or horticultural host plant species.

Genomic DNA prepared from the source plant is fragmented (e.g. by DNaseI, restriction enzymes, or mechanically) and cloned into a vector suitable for making plant genomic libraries, such as pGA482 (An. G., 1995, Methods Mol. Biol. 44:47-58). This vector contains the *A. tumefaciens* left and right borders needed for gene transfer to plant cells and antibiotic markers for selection in *E. coli*, *Agrobacterium*, and plant cells. A multicloning site is provided for insertion of the genomic, fragments. A cos sequence is present for the efficient packaging of DNA into bacteriophage lambda heads for transfection of the primary library into *E. coli*. The vector accepts DNA fragments of 25-40 kb.

The primary library can also be directly electroporated into an *A. tumefaciens* or *A. rhizogenes* strain that is used to infect and transform host plant cells (Main, GD et al., 1995, Methods Mol. Biol. 44:405-412). Alternatively, DNA can be introduced by electroporation or PEG-mediated uptake into protoplasts of the recipient plant species (Bilang et al. (1994) Plant Mol. Biol Manual, Kluwer Academic Publishers, Al: 1- 16) or by particle bombardment of cells or tissues (Christou, ibid, A2:1-15). If necessary, antibiotic markers in the T-DNA region can be eliminated, as long as selection for the trait is possible, so that the final plant products contain no antibiotic genes.

Stably transformed whole cells acquiring the trait are selected on solid or liquid media containing the agent to which the introduced DNA confers resistance or tolerance. If the trait in question cannot be selected for directly, transformed cells can be selected with antibiotics and allowed to form callus or regenerated to whole plants and then screened for the desired property.

The second and further cycles consist of isolating genomic DNA from each transgenic line and introducing it into one or more of the other transgenic lines. In each round, transformed cells are selected or screened for incremental improvement. To speed the process of using multiple cycles of transformation, plant regeneration can be deferred until the last round. Callus tissue generated from the protoplasts or transformed tissues can serve as a source of genomic DNA and new host cells. After the final round, fertile plants are regenerated and the progeny are selected for homozygosity of the inserted DNAs. Ultimately, a new plant is created that carries multiple inserts which additively or synergistically combine to confer high levels of the desired trait. Alternatively, microspores can be isolated as homozygotes generated from spontaneous diploids.

In addition, the introduced DNA that confers the desired trait can be traced because it is flanked by known sequences in the vector. Either PCR or plasmid rescue is used to isolate the sequences and characterize them in more detail. Long PCR (Foord, OS and Rose, EA, 1995, PCR Primer: A Laboratory Manual , CSBL Press, pp 63-77) of the full 25-40 kb insert is achieved with the proper reagents and techniques using as primers the T-DNA border sequences. If the vector is modified to contain the *E. coli* origin of replication and an antibiotic marker between the T-DNA borders, a rare cutting restriction enzyme, such as NotI or SfiI, that cuts only at the ends of the inserted DNA is used to create fragments containing the source plant DNA that are then self-ligated and transformed into *E. coli* where they replicate as plasmids. The total DNA or subfragment of it that is responsible for the transferred trait can be subjected to in vitro evolution by DNA stochastic &/or non-stochastic mutagenesis. The stochastic &/or non-stochastic mutagenized library can be reiteratively recombined by any method herein and then introduced into host plant cells and screened for improvement of the trait. In this way, single and multigene traits can be transferred from one species to another and optimized for higher expression or activity leading to whole organism improvement. This entire process can also be reiteratively repeated. Alternatively, the cells can be transformed microspores with the regenerated haploid plants being screened directly for improved traits as noted below.

4.12.10 PLANT CELL IS PUT INTO CONTACT WITH AGENT TO SEE WHICH CELLS SURVIVE.

After a suitable period of incubation to allow recombination to occur and for expression of recombinant genes, the plant cells are contacted with the agent to which resistance is to be acquired, and surviving plant cells are collected. Some or all of these plant cells can be subject to a further round of recombination and screening. Eventually, plant cells having the required degree of resistance are obtained.

These cells can then be cultured into transgenic plants. Plant regeneration from cultured protoplasts is described in Evans et al., "Protoplast Isolation and Culture," Handbook of Plant Cell Cultures 1, 124-176 (MacMillan Publishing Co., New York, 1983); Davey, "Recent Developments in the Culture and Regeneration of Plant Protoplasts," Protoplasts, (1983) pp. 12-29, (Birkhauser, Basal 1983); Dale, "Protoplast Culture and Plant Regeneration of Cereals and Other Recalcitrant Crops," Protoplasts (1983) pp. 31-41, (Birkhauser, Basel 1983); Binding, "Regeneration of Plants," Plant Protoplasts, pp. 21-73, (CRC Press, Boca Raton, 1985).

4.12.11 START IN BACTERIAL CELL SINCE FASTER EVOLUTION AND TRANSFORM INTO PLANT

In a variation of the above method, one or more preliminary rounds of recombination and screening can be performed in bacterial cells according to the same general strategy as described for plant cells. More rapid evolution can be achieved in bacterial cells due to their greater growth rate and the greater efficiency with which DNA can be introduced into such cells. After one or more rounds of recombination/screening, a DNA fragment library is recovered from bacteria and transformed into the plant cells. The library can either be a complete library or a focused library. A focused library can be produced by amplification from primers specific for plant sequences, particularly plant sequences known or suspected to have a role in conferring resistance.

4.12.12 MICROSPORE MANIPULATION

Microspores are haploid ($1n$) male spores that develop into pollen grains. Anthers contain a large numbers of microspores in early-uninucleate to first-mitosis stages. Microspores have been successfully induced to develop into plants for most species, such as, e.g., rice (Chen, CC 1977 *In Vitro.* 13: 484-489), tobacco (Atanassov, I. et al. 1998 *Plant Mol Biol.* 38:1169-1178), Tradescantia (Savage JRK and Papworth DG. 1998 *Mutat Res.* 422:313-322), Arabidopsis (Park SK et al. 1998 Development. 125:3789- 3799), sugar beet (Majewska-Sawka A and Rodrigues-Garcia NE 1996 *J Cell Sci.* 109:859-866), Barley (Olsen FL 1991 *Hereditas* 115:255-266) and oilseed rape (Boutillier KA et al. 1994 *Plant Mol Biol.* 26:1711-1723).

The plants derived from microspores are predominantly haploid or diploid (infrequently polyploid and aneuploid). The diploid plants are homozygous and fertile and can be generated in a relatively short time. Microspores obtained from F1 hybrid plants represent great diversity, thus being an excellent model for studying recombination. In addition, microspores can be transformed with T-DNA introduced by agrobacterium or other available means and then regenerated into individual plants. Furthermore, protoplasts can be made from microspores and they can be fused similar to what occur in fungi and bacteria.

Microspores, due to their complex ploidy and regenerating ability, provide a tool for plant whole genome stochastic &/or non-stochastic mutagenesis. For example, if pollens from 4 parents are collected 4 and pooled, and then used to randomly pollinate the parents, the progenies should have $2^4 = 16$ possible combinations. Assuming this plant has 7 chromosomes, microspores collected from the 16 progenies will represent $2^7 \times 16 = 2048$ possible chromosomal combinations. This number is even greater if meiotic processes occur. When diploid, homozygous embryos are generated from these microspores, in many cases, they are screened for desired phenotypes, such as herbicide- or disease- resistant. In addition, for plant oil composition these embryos can be dissected into two halves: one for analysis the other for regeneration into a

viable plant. Protoplasts generated from microspores (especially the haploid ones) are pooled and fused. Microspores obtained from plants generated by protoplast fusion are pooled and fused again, increasing the genetic diversity of the resulting microspores. Microspores can be subjected to mutagenesis in various ways, such as by chemical mutagenesis, radiation-induced mutagenesis and, e.g., t-DNA transformation, prior to fusion or regeneration. New mutations which are generated can be recombined through the recursive processes described above and herein.

4.12.13 ACQUISTION OF SALT TOLERANCE

DNA from a salt tolerant plant is isolated and used to create a genomic library. Protoplasts made from the recipient species are transformed/transfected with the genomic library (e.g., by electroporation, agrobacterium, etc.). Cells are selected on media with a normally inhibitory level of NaCl. Only the cells with newly acquired salt tolerance will grow into callus tissue. The best lines are chosen and genomic libraries are made from their pooled DNA. These libraries are transformed into protoplasts made from the first round transformed calli. Again, cells are selected on increased salt concentrations. After the desired level of salt tolerance is achieved, the callus tissue can be induced to regenerate whole plants. Progeny of these plants are typically analyzed for homozygosity of the inserts to ensure stability of the acquired trait. At the indicated steps, plant regeneration or isolation and stochastic &/or non-stochastic mutagenesis of the introduced genes can be added to the overall protocol.

4.13 EVOLVE TRANSGENIC ANIMALS

4.13.1 OPTIMIZE TRANSGENE

One goal of transgenesis is to produce transgenic animals, such as mice, rabbits, sheep, pigs, goats, and cattle, secreting a recombinant protein in the milk. A transgene for this purpose typically comprises in operable linkage a promoter and an enhancer from a milk-protein gene (e.g., alpha, beta, or gamma casein, beta-lactoglobulin, acid whey protein or alpha-lactalbumin), a signal sequence, a recombinant protein coding

sequence and a transcription termination site.

Optionally, a transgene can encode multiple chains of a multichain protein, such as an immunoglobulin, in which case, the two chains are usually individually operably linked to sets of regulatory sequences. Transgenes can be optimized for expression and secretion by recursive sequence recombination. Suitable substrates for recombination include regulatory sequences such as promoters and enhancers from milk-protein genes from different species or individual animals. Cycles of recombination can be performed in vitro or in vivo by any of the formats discussed. Screening is performed in vivo on cultures of mammary-gland derived cells, such as HC11 or MacT, transfected with transgenes and reporter constructs such as those discussed above. After several cycles of recombination and screening, transgenes resulting in the highest levels of expression and secretion are extracted from the mammary gland tissue culture cells and used to transfect embryonic cells, such as zygotes and embryonic stem cells, which are matured into transgenic animals.

4.13.2 OPTIMIZE WHOLE ANIMAL BY TRANSFORMING INTO EMBRYONIC CELLS GENE OF DESIRED TRAIT

4.13.2.1 GROWTH HORMONE

In this approach, libraries of incoming fragments are transformed into embryonic cells, such as ES cells or zygotes. The fragments can be variants of a gene known to confer a desired property, such as growth hormone. Alternatively, the fragments can be partial or complete genomic libraries including many genes. Fragments are usually introduced into zygotes by microinjection as described by Gordon et al., Methods Enzymol. 10 1, 414 (1984); Hogan et al., Manipulation of the Mouse Embryo: A Laboratory Manual (C.S.H.L. N.Y., 1986) (mouse embryo), and Hammer et al., Nature 315, 680 (1985) (rabbit and porcine embryos); Gandolfi et al., J Reprod. Fert. 81, 23-28 (1987); Rexroad et al., J Anim. Sci. 66, 947-953 (1988) (ovine embryos) and Eyestone et al., J Reprod. Fert. 85, 715-720 (1989); Camous et al., J Reprod. Fert. 72, 779-785 (1984); and Heyman et al., Theriogenology 27, 5968 (1987) (bovine embryos). Zygotes are then matured and introduced into recipient female animals

which gestate the embryo and give birth to a transgenic offspring.

Alternatively, transgenes can be introduced into embryonic stem cells (ES).

These cells are obtained from preimplantation embryos cultured in vitro. Bradley et al., *Nature* 309, 255-258 (1984). Transgenes can be introduced into such cells by electroporation or microinjection. Transformed ES cells are combined with blastocysts from a non-human animal. The ES cells colonize the embryo and in some embryos form the germ line of the resulting chimeric animal. See Jaenisch, *Science*, 240, 1468-1474 (1988).

Regardless whether zygotes or ES are used, screening is performed on whole animals for a desired property, such as increased size and/or growth rate. DNA is extracted from animals having evolved toward acquisition of the desired property. This DNA is then used to transfet further embryonic cells. These cells can also be obtained from animals that have acquired toward the desired property in a split and pool approach. That is, DNA from one subset of such animals is transformed into embryonic cells prepared from another subset of the animals. Alternatively, the DNA from animals that have evolved toward acquisition of the desired property can be transfected into fresh embryonic cells. In either alternative, transfected cells are matured into transgenic animals, and the animals subjected to a further round of screening for the desired property.

Initially, a library is prepared of variants of a growth hormone gene. The variants can be natural or induced. The library is coated with recA protein and transfected into fertilized fish eggs. The fish eggs then mature into fish of different sizes. The growth hormone gene fragment of genomic DNA from large fish is then amplified by PCR and used in the next round of recombination. Alternatively, fish -IFN is evolved to enhance resistance to viral infections as described below.

4.13.2.2 EVOLUTION OF IMPROVED HORMONES FOR EXPRESSION IN TRANSGENIC ANIMALS

4.13.3 TO CREATE ANIMALS WITH IMPROVED TRAITS

Evolution of improved hormones for expression in transgenic animals (e.g., Fish) to create animals with improved traits. Hormones and cytokines are key regulators of size, body weight, viral resistance and many other commercially important traits. DNA stochastic &/or non-stochastic mutagenesis is used to rapidly evolve the genes for these proteins using *in vitro* assays. This was demonstrated with the evolution of the human alpha interferon genes to have potent antiviral activity on murine cells. Large improvements in activity were achieved in two cycles of family stochastic &/or non-stochastic mutagenesis of the human IFN genes.

In general, a method of increasing resistance to virus infection in cells can be performed by first introducing a stochastic &/or non-stochastic mutagenized library comprising at least one stochastic &/or non-stochastic mutagenized interferon gene into animal cells to create an initial library of animal cells or animals. The initial library is then challenged with the virus. Animal cells or animals are selected from the initial library which are resistant to the virus and a plurality of transgenes from a plurality of animal cells or animals which are resistant to the virus are recovered. The plurality of transgenes is recovered to produce an evolved library of animal cells or animals which is again challenged with the virus. Cells or animals are selected from the evolved library which are resistant to the virus.

For example, genes evolved with *in vitro* assays are introduced into the germplasm of animals or plants to create improved strains. One limitation of this procedure is that *in vitro* assays are often only crude predictors of *in vivo* activity. However, with improving methods for the production of transgenic plants and animals, one can now marry whole organism breeding with molecular breeding. The approach is to introduce stochastic &/or non-stochastic mutagenized libraries of hormone genes into the species of interest. This can be done with a single gene per transgenic or with pools of genes per transgenic. Progeny are then screened for the phenotype of interest. In this case, stochastic &/or non-stochastic mutagenized libraries of interferon genes

(alpha IFN for example) are introduced into transgenic fish. The library of transgenic fish are challenged with a virus. The most resistant fish are identified (i.e. either survivors of a lethal challenge; or those that are deemed most 'healthy' after the challenge). The IFN transgenes are recovered by PCR and stochastic &/or non-stochastic mutagenized in either a poolwise or a pairwise fashion. This generates an evolved library of IFN genes. A second library of transgenic fish is created and the process is repeated. In this way, IFN is evolved for improved antiviral activity in a whole organism assay. This procedure is general and can be applied to any trait that is affected by a gene or gene family of interest and which can be quantitatively measured.

Fish interferon sequence data is available for the Japanese flatfish (*Paralichthys olivaceus*) as mRNA sequence (Tamai et al (1993) "Cloning and expression of flatfish (*Paralichthys olivaceus*) interferon cDNA." *Biochem. Biophys. Acta* 1174, 182-186; Y see also, Tami et al. (1993) "Purification and characterization of interferon-like antiviral protein derived from flatfish (*Paralichthys olivaceus*) lymphocytes immortalized by oncogenes." *Cytotechnology* 1993; 11 (2):121-131). This sequence can be used to clone out IFN genes from this species. This sequence can also be used as a probe to clone homologous interferons from additional species of fish. As well, additional sequence information can be utilized to clone out more species of fish interferons. Once a library of interferons has been cloned, these can be family stochastic &/or non-stochastic mutagenized to generate a library of variants.

In one embodiment, BHK-21 (A fibroblast cell line from hamster) can be transfected with the stochastic &/or non-stochastic mutagenized WN-expression plasmids. Active recombinant IFN is produced and then purified by WGA agarose affinity chromatography (Tamai, et al. 1993 *Biochim Ciophys Acta. supra*). The antiviral activity of IFN can be measured on fish cells challenged by rhabdovirus. Tami et al. (1993) "Purification and characterization of interferon-like antiviral protein derived from flatfish (*Paralichthys olivaceus*) lymphocytes immortalized by oncogenes. "

Cytotechnolgy 1993; 11 (2):121-131).

4.13.4 WHOLE GENOME STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS IN HIGHER ORGANISMS POOLWISE RECURSIVE BREEDING

The present invention provides a procedure for generating large combinatorial libraries of higher eukaryotes, plants, fish, domesticated animals, etc. In addition to the procedures outlined above, poolwise combination of male and female gametes can also be used to generate large diverse molecular libraries.

In one aspect, the process includes recursive poolwise matings for several generations without any deliberate screening. This is similar to classical breeding, except that pools of organisms, rather than pairs of organisms, are mated, thereby accelerating the generation of genetic diversity. This method is similar to recursive fusion of a diverse population of bacterial protoplasts resulting in the generation of multiparent progeny harboring genetic information from all of the starting population of bacteria. The process described here is to perform analogous artificial or natural matings of large populations of natural isolates, imparting a split pool mating strategy. Before mating, all of the male gametes i.e. pollen, sperm, etc., are isolated from the starting population and pooled. These are then used to "self" fertilize a mixed pool of the female gametes from the same population.

The process is repeated with the subsequent progeny for several generations, with the final progeny being a combinatorial organism library with each member having genetic information originating from many if not all of the starting "parents." This process generates large diverse organism libraries on which many selections and or screens can be imparted, and it does not require sophisticated in vitro manipulation of genes. However, it results in the creation of useful new strains (perhaps well diluted in the population) in a much shorter time frame than such organisms could be generated

using a classical targeted breeding approach.

These libraries are generated relatively quickly (e.g., typically in less than three years for most plants of commercial interest, with six cycles or less of recursive breeding being sufficient to generate desired diversity). An additional benefit of these methods is that the resulting libraries provide organismal diversity in areas, such as agriculture, aquaculture, and animal husbandry, that are currently genetically homogeneous.

Examples of these methods for several organisms are described below.

4.13.5 PLANTS

Plants A population of plants, for example all of the different corn strains in a commercial seed/germplasm collection, are grown and the pollen from the entire population is harvested and pooled. This mixed pollen population is then used to "self" fertilize the same population. Self pollination is prevented, so that the fertilization is combinatorial. The cross results in all pairwise crosses possible within the population, and the resulting seeds result in many of the possible outcomes of each of these pairwise crosses. The seeds from the fertilized plants are then harvested, pooled, planted, and the pollen is again harvested, pooled, and used to "self" fertilize the population. After only several generations, the resulting population is a very diverse combinatorial library of corn. The seeds from this library are harvested and screened for desirable traits, e.g., salt tolerance, growth rate, productivity, yield, disease resistance, etc. Essentially any plant collection can be modified by this approach. Important commercial crops include both monocots and dicots. Monocots include plants in the grass family (Gramineae), such as plants in the sub families Fetucoideae and Poacoideae, which together include several hundred genera including plants in the genera Agrostis, Phleum, Dactylis, Sorgum, Setaria, Zea (e.g., corn), Oryza (e.g., rice), Triticum (e.g., wheat), Secale (e.g., rye), Avena (e.g., oats), Hordeum (e.g., barley), Saccharum, Poa, Festuca, Stenotaphrum, Cynodon, Coix, the Olyreae, Phareae and many others. Plants in the family Gramineae are a particularly

preferred target plants for the methods of the invention.

Additional preferred targets include other commercially important crops, e.g., from the families Compositae (the largest family of vascular plants, including at least 1,000 genera, including important commercial crops such as sunflower), and Leguminosae or "pea family," which includes several hundred genera, including many commercially valuable crops such as pea, beans, lentil, peanut, yam bean, cowpeas, velvet beans, soybean, clover, alfalfa, lupine, vetch, lotus, sweet clover, wisteria, and sweetpea. Common crops applicable to the methods of the invention include Zea mays, rice, soybean, sorghum, wheat, oats, barley, millet, sunflower, and canola.

This process can also be carried out using pollen from different species or more divergent strains (e.g., crossing the ancient grasses with corn). Different plant species can be forced to cross. Only a few plants from an initial cross would have to result in order to make the process viable. These few progeny, e.g., from a cross between soy bean and corn, would generate pollen and eggs, each of which would represent a different meiotic outcome from the recombination of the two genomes. The pollen would be harvested and used to "self pollinate the original progeny. This process would then be carried out recursively. This would generate a large family stochastic &/or non-stochastic mutagenized library of two or more species, which could be subsequently screened.

4.13.6 FISH

Fish The natural tendency of fish to lay their eggs outside of the body and to have a male cover those eggs with sperm provides another opportunity for a split pooled breeding strategy. The eggs from many different fish, e.g., salmon from different fisheries about the world, can be harvested, pooled, and then fertilized with similarly collected and pooled salmon sperm. The fertilization will result in all of the possible pairwise matings of the starting population. The resulting progeny is then grown and again the sperm and eggs are harvested, and pooled, with each egg and sperm

representing a different meiotic outcome of the different crosses. The pooled sperm are then used to fertilize the pooled eggs and the process is carried out recursively. After several generations the resulting progeny can then be subjected to selections and screens for desired properties, such as size, disease resistance, etc.

4.13.7 ANIMALS

Animals The advent of in vitro fertilization and surrogate motherhood provides a means of whole genome stochastic &/or non-stochastic mutagenesis in animals such as mammals. As with fish, the eggs and the sperm from a population, for example from all slaughter cows, are collected and pooled. The pooled eggs are then in vitro fertilized with the pooled sperm. The resulting embryos are then returned to surrogate mothers for development. As above, this process is repeated recursively until a large diverse population is generated that can be screened for desirable traits.

A technically feasible approach would be similar to that used for plants. In this case, sperm from the males of the starting population is collected and pooled, and then this pooled sample is used to artificially inseminate multiple females from each of the starting populations. Only one (or a few) sperm would succeed in each animal, but these should be different for each fertilization. The process is reiterated by harvesting the sperm from all of the male progeny, pooling it, and using it to fertilize all of the female progeny. The process is carried out recursively for several generations to generate the organism library, which can then be screened.

4.14 PREDICTIVE TOOL IN LOOKING FOR DRUGS

Recursive sequence recombination can be used to simulate natural evolution of pathogenic microorganisms in response to exposure to a drug under test. Using recursive sequence recombination, evolution proceeds at a faster rate than in natural evolution. One measure of the rate of evolution is the number of cycles of recombination and screening required until the microorganism acquires a defined level of resistance to the drug. The information from this analysis is of value in

comparing the relative merits of different drugs and in particular, in predicting their long term efficacy on repeated administration.

The pathogenic microorganisms used in this analysis include the bacteria that are a common source of human infections, such as chlamydia, rickettsial bacteria, mycobacteria, staphylococci, streptococci, pneumonococci, meningococci and congoocci, klebsiella, proteus, serratia, pseudomonas, legionella, diphtheria, salmonella, bacilli, cholera, tetanus, botulism, anthrax, plague, leptospirosis, and Lymes disease bacteria.

Evolution is effected by transforming an isolate of bacteria that is sensitive to a drug under test with a library of DNA fragments. The fragments can be a mutated version of the genome of the bacteria being evolved. If the target of the drug is a known protein or nucleic acid, a focused library containing variants of the corresponding gene can be used. Alternatively, the library can come from other kinds of bacteria, especially bacteria typically found inhabiting human tissues, thereby simulating the source material available for recombination in vivo. The library can also come from bacteria known to be resistant to the drug. After transformation and propagation of bacteria for an appropriate period to allow for recombination to occur and recombinant genes to be expressed, the bacteria are screened by exposing them to the drug under test and then collecting survivors. Surviving bacteria are subject to further rounds of recombination. The subsequent round can be effected by a split and pool approach in which DNA from one subset of surviving bacteria is introduced into a second subset of bacteria. Alternatively, a fresh library of DNA fragments can be introduced into surviving bacteria. Subsequent round(s) of selection can be performed at increasing concentrations of drug, thereby increasing the stringency of selection.

4.14.1 BIOSYNTHESIS

Metabolic engineering can be used to alter organisms to optimize the production of

practically any metabolic intermediate, including antibiotics, vitamins, amino acids such as phenylalanine and aromatic amino acids, ethanol, butanol, polymers such as xanthan gum and bacterial cellulose, peptides, and lipids. When such compounds are already produced by a host, the recursive sequence recombination techniques described above can be used to optimize production of the desired metabolic intermediate, including such features as increasing enzyme substrate specificity and turnover number, altering metabolic fluxes to reduce the concentrations of toxic substrates or intermediates, increasing resistance of the host to such toxic compounds, eliminating, reducing or altering the need for inducers of gene expression/activity, increasing the production of enzymes necessary for metabolism, etc.

Enzymes can also be evolved for improved activity in solvents other than water. This is useful because intermediates in chemical syntheses are often protected by blocking groups which dramatically affect the solubility of the compound in aqueous solvents. Many compounds can be produced by a combination of pure chemical and enzymically catalyzed reactions. Performing enzymic reactions on almost insoluble substrates is clearly very inefficient, so the availability of enzymes that are active in other solvents will be of great use. One example of such a scheme is the evolution of a para-nitrobenzyl esterase to remove protecting groups from an intermediate in loracarbef synthesis (Moore, J.C. and Arnold, F.H. *Nature Biotechnology* 14:458-467 (1996)). In this case alternating rounds of error-prone PCR and colony screening for production of a fluorescent reporter from a substrate analogue were used to generate a mutant esterase that was 16-fold more active than the parent molecule in 30% dimethylformamide. No individual mutation was found to contribute more than a 2-fold increase in activity, but it was the combination of a number of mutations which led to the overall increase.

Structural analysis of the mutant protein showed that the amino acid changes were distributed throughout the length of the protein in a manner that could not have been rationally predicted. Sequential rounds of error-prone PCR have the problem that after each round all but one mutant is discarded, with a concomitant loss of information contained in all the other beneficial mutations. Recursive sequence recombination avoids this problem, and would thus be ideally suited to evolving enzymes for

catalysis in other solvents, as well as in conditions where salt concentrations or pH were different from the original enzyme optimas.

In addition, the yield of almost any metabolic pathway can be increased, whether consisting entirely of genes endogenous to the host organisms or all or partly heterologous genes. Optimization of the expression levels of the enzymes in a pathway is more complex than simply maximizing expression. In some cases regulation, rather than constitutive expression of an enzyme may be advantageous for cell growth and therefore for product yield, as seen for production of phenylalanine (Backman et al. Ann. NY Acad. Sci. 589:16-24 (1990)) and 2-keto-L- gluconic acid (Anderson et al. U.S. 5,032,514). In addition, it is often advantageous for industrial purposes to express proteins in organisms other than their original hosts. New host strains may be preferable for a variety of reasons, including ease of cloning and transformation, pathogenicity, ability to survive in particular environments and a knowledge of the physiology and genetics of the organisms. However, proteins expressed in heterologous organisms often show markedly reduced activity for a variety of reasons including inability to fold properly in the new host (Sarthy et al. Appl. Environ. Micro. 53:1996-2000 (1987)). Such difficulties can indeed be overcome by the recursive sequence recombination strategies of the instant invention.

4.14.2 ANTIBIOTICS

The range of natural small molecule antibiotics includes but is not limited to peptides, peptidolactones, thiopeptides, beta-lactams, glycopeptides, lantibiotics, microcins, polyketide-derived antibiotics (anthracyclines, tetracyclines, macrolides, avermectins, polyethers and ansamycins), chloramphenicol, aminoglycosides, aminocyclitols, polyoxins, agrocin and isoprenoids. There are at least three ways in which recursive sequence recombination techniques of the instant invention can be used to facilitate novel drug synthesis, or to improve biosynthesis of existing antibiotics.

First, antibiotic synthesis enzymes can be "evolved" together with transport systems that allow entry of compounds used as antibiotic precursors to improve uptake and incorporation of function-altering artificial side chain precursors. For example,

penicillin V is produced by feeding *Penicillium* the artificial side chain precursor phenoxyacetic acid, and LY146032 by feeding *Streptomyces roseosporus* decanoic acid (Hopwood, Phil. Trans. R. Soc. Lond. B 324:549-562 (1989)). Poor precursor uptake and poor incorporation by the synthesizing enzyme often lead to inefficient formation of the desired product. Recursive sequence recombination of these two systems can increase the yield of desired product.

Furthermore, a combinatorial approach can be taken in which an enzyme is stochastic &/or non-stochastic mutagenized for novel catalytic activity/substrate recognition (perhaps by including randomizing oligonucleotides in key positions such as the active site). A number of different substrates (for example, analogues of side chains that are normally incorporated into the antibiotic) can then be tested in combination with all the different enzymes and tested for biological activity. In this embodiment, plates are made containing different potential antibiotic precursors (such as the side chain analogues). The microorganisms containing the stochastic &/or non-stochastic mutagenized library (the library strain) are replicated onto those plates, together with a competing, antibiotic sensitive, microorganism (the indicator strain). Library cells that are able to incorporate the new side chain to produce an effective antibiotic will thus be able to compete with the indicator strain, and will be selected for.

Second, the expression of heterologous genes transferred from one antibiotic synthesizing organism to another can be optimized. The newly introduced enzyme(s) act on secondary metabolites in the host cell, transforming them into new compounds with novel properties. Using traditional methods, introduction of foreign genes into antibiotic synthesizing hosts has already resulted in the production of novel hybrid antibiotics. Examples include mederhodin, dihydrogranatirhodin, 6-deoxyerythromycin A, isovalerylspiramycin and other hybrid macrolides (Cameron et. al. Appl. Biochem. Biotechnol. 38:105-140 (1993)). The recursive sequence recombination techniques of the instant invention can be used to optimize expression of the foreign genes, to stabilize the enzyme in the new host cell, and to increase the activity of the introduced enzyme against its new substrates in the new host cell. In some embodiments of the invention, the host genome may also be so optimized.

Third, the substrate specificity of an enzyme involved in secondary metabolism can

be altered so that it will act on and modify a new compound or so that its so activity is changed and it acts at a different subset of positions of its normal substrate. Recursive sequence recombination can be used to alter the substrate specificities of enzymes. Furthermore, in addition to recursive sequence recombination of individual enzymes being a strategy to generate novel antibiotics, recursive sequence recombination of entire pathways, by altering enzyme ratios, will alter metabolite fluxes and may result, not only in increased antibiotic synthesis, but also in the synthesis of different antibiotics. This can be deduced from the observation that expression of different genes from the same cluster in a foreign host leads to different products being formed (see p. 80 in Hutchinson et. al., (1991) Ann NY Acad Sci, 646:78-93).

Recursive sequence recombination of the introduced gene clusters may result in a variety of expression levels of different proteins within the cluster (because it produces different combinations of, in this case regulatory, mutations). This in turn may lead to a variety of different end products. Thus, "evolution" of an existing antibiotic synthesizing pathway could be used to generate novel antibiotics either by modifying the rates or substrate specificities of enzymes in that pathway.

Additionally, antibiotics can also be produced in vitro by the action of a purified enzyme on a precursor. For example isopenicillin N synthase catalyses the cyclization of many analogues of its normal substrate (d-(L-a-amino adipyl)-L-cysteinyl-D-valine) (Hutchinson, Med. Res. Rev. 8:557-567 (1988)). Many of these products are active as antibiotics. A wide variety of substrate analogues can be tested for incorporation by secondary metabolite synthesizing enzymes without concern for the initial efficiency of the reaction. Recursive sequence recombination can be used subsequently to increase the rate of reaction with a promising new substrate.

Thus, organisms already producing a desired antibiotic can be evolved with the recursive sequence recombination techniques described above to maximize production of that antibiotic. Additionally, new antibiotics can be evolved by manipulation of genetic material from the host by the recursive sequence recombination techniques described above. Genes for antibiotic production can be

transferred to a preferred host after cycles of recursive sequence recombination or can be evolved in the preferred host as described above.

Antibiotic genes are generally clustered and are often positively regulated, making them especially attractive candidates for the recursive sequence recombination techniques of the instant invention. Additionally, some genes of related pathways show cross-hybridization, making them preferred candidates for the generation of new pathways for new antibiotics by the recursive sequence recombination techniques of the invention.

Furthermore, increases in secondary metabolite production including enhancement of substrate fluxes (by increasing the rate of a rate limiting enzyme, deregulation of the pathway by suppression of negative control elements or over expression of activators and the relief of feedback controls by mutation of the regulated enzyme to a feedback-insensitive deregulated protein) can be achieved by recursive sequence recombination without exhaustive analysis of the regulatory mechanisms governing expression of the relevant gene clusters.

The host chosen for expression of evolved genes is preferably resistant to the antibiotic produced, although in some instances production methods can be designed so as to sacrifice host cells when the amount of antibiotic produced is commercially significant yet lethal to the host. Similarly, bioreactors can be designed so that the growth medium is continually replenished, thereby "drawing off" antibiotic produced and sparing the lives of the producing cells. Preferably, the mechanism of resistance is not the degradation of the antibiotic produced.

Numerous screening methods for increased antibiotic expression are known in the art, as discussed above, including screening for organisms that are more resistant to the antibiotic that they produce. This may result from linkage between expression of the antibiotic synthesis and antibiotic resistance genes (Chater, Bio/Technology 8:115-121 (1990)). Another screening method is to fuse a reporter gene (e.g. *xylE* from the

Pseudomonas TOL plasmid) to the antibiotic production genes. Antibiotic synthesis gene expression can then be measured by looking for expression of the reporter (e.g. xylE encodes a catechol dioxygenase which produces yellow muconic semialdehyde when colonies are sprayed with catechol (Zukowski et al. Proc. Natl. Acad. Sci. U.S.A. 80:1101-1105 (1983)). The wide variety of cloned antibiotic genes provides a wealth of starting materials for the recursive sequence recombination techniques of the instant invention. For example, genes have been cloned from Streptomyces cattleya which direct cephalexin C synthesis in the non-antibiotic producer Streptomyces lividans (Chen et al. Bio/Technology 6:1222-1224 (1988)). Clustered genes for penicillin biosynthesis (-L--aminoacidyl)-L-cysteinyl-D-valine synthetase; isopenicillin N synthetase and acyl coenzyme A:6-aminopenicillanic acid acyltransferase) have been cloned from Penicillium chrysogenum. Transfer of these genes into Neurospora crassa and Aspergillus niger result in the synthesis of active penicillin V (Smith et al. Bio/Technology 8:39-41 (1990)). For a review of cloned genes involved in Cephalosporin C, Penicillins G and V and Cephalexin C biosynthesis, see Piepersberg, Crit. Rev. Biotechnol. 14:251-285 (1994). For a review of cloned clusters of antibiotic-producing genes, see Chater Bio/Technology 8:115-121 (1990). Other examples of antibiotic synthesis genes transferred to industrial producing strains, or over expression of genes, include tylosin, cephalexin C, cephalosporin C, LL-E33288 complex (an antitumor and antibacterial agent), doxorubicin, spiramycin and other macrolide antibiotics, reviewed in Cameron et al. Appl. Biochem. Biotechnol. 38:105-140 (1993).

4.14.3 BIOSYNTHESIS TO REPLACE CHEMICAL SYNTHESIS OF ANTIBIOTICS

Some antibiotics are currently made by chemical modifications of biologically produced starting compounds. Complete biosynthesis of the desired molecules may currently be impractical because of the lack of an enzyme with the required enzymatic activity and substrate specificity. For example, 7-aminodeacetoxycephalosporanic acid (7-ADCA) is a precursor for semi-synthetically produced cephalosporins. 7-ADCA is made by a chemical ring expansion from penicillin V followed by

enzymatic deacylation of the phenoxyacetal group. Cephalosporins could in principle be produced biologically from their corresponding penicillins (e.g., cephalosporin V or G from penicillin V or G) using penicillin N expandase, but other penicillins (such as penicillin V or G) are not used as substrates by known expandases. The recursive sequence recombination techniques of the invention can be used to alter the enzyme so that it will use penicillin V as a substrate. Similarly, penicillin transacylase could be so modified to accept cephalosporins or cephemycins as substrates.

In yet another example, penicillin amidase expressed in *E. coli* is a key enzyme in the production of penicillin G derivatives. The enzyme is generated from a precursor peptide and tends to accumulate as insoluble aggregates in the periplasm unless non-metabolizable sugars are present in the medium (Scherrer et al., *Appl. Microbiol. Biotechnol.* 42:85-91 (1994)). Evolution of this enzyme through the methods of the instant invention could be used to generate an enzyme that folds better, leading to a higher level of active enzyme expression.

In yet another example, Penicillin G acylase covalently linked to agarose is used in the synthesis of penicillin G derivatives. The enzyme can be stabilized for increased activity, longevity and/or thermal stability by chemical modification (Fernandez-Lafuente et. al. *Enzyme Microb. Technol.* 14:489-495 (1992)). Increased thermal stability is an especially attractive application of the recursive sequence recombination techniques of the instant invention, which can obviate the need for the chemical modification of such enzymes. Selection for thermostability can be performed *in vivo* in *E. coli* or in thermophiles at higher temperatures. In general, thermostability is a good first step in enhancing general stabilization of enzymes. Random mutagenesis and selection can also be used to adapt enzymes to function in non-aqueous solvents (Arnold *Curr Opin Biotechnol.* 4:450-455 (1993); Chen et. al. *Proc. Natl. Acad. Sci. U.S.A.*, 90:5618-5622 (1993)). Recursive sequence recombination represents a more powerful (since recombinogenic) method of generating mutant enzymes that are stable and active in non-aqueous environments. Additional screening can be done on the basis of enzyme stability in solvents.

4.14.4 POLYKETIDES

Polyketides include antibiotics such as tetracycline and erythromycin, anti-cancer agents such as daunomycin, immunosuppressants such as FK506 and rapamycin and veterinary products such as monesin and avermectin. Polyketide synthases (PKS's) are multifunctional enzymes that control the chain length, choice of chain-building units and reductive cycle that generates the huge variation in naturally occurring polyketides. Polyketides are built up by sequential transfers of "extender units" (fatty acyl CoA groups) onto the appropriate starter unit (examples are acetate, coumarate, propionate and malonamide). The PKS's determine the number of condensation reactions and the type of extender groups added and may also fold and cyclize the polyketide precursor. PKS's reduce specific -keto groups and may dehydrate the resultant -hydroxyls to form double bonds. Modifications of the nature or number of building blocks used, positions at which -keto groups are reduced, the extent of reduction and different positions of possible cyclizations, result in formation of different final products. Polyketide research is currently focused on modification and inhibitor studies, site directed mutagenesis and 3-D structure elucidation to lay the groundwork for rational changes in enzymes that will lead to new polyketide products.

Recently, McDaniel et al. (Science 262:1546- 1550 (1995)) have developed a Streptomyces host-vector system for efficient construction and expression of recombinant PKSs. Hutchinson (Bio/Technolo 12:375-308 (1994)) reviewed targeted mutation of specific biosynthetic genes and suggested that microbial isolates can be screened by DNA hybridization for genes associated with known pharmacologically active agents so as to provide new metabolites or increased yields of metabolites already being produced. In particular, that review focuses on polyketide synthase and pathways to aminoglycoside and oligopeptide antibiotics.

The recursive sequence recombination techniques of the instant invention can be used to generate modified enzymes that produce novel polyketides without such detailed analytical effort. The availability of the PKS genes on plasmids and the existence of

E. coli- *Streptomyces* shuttle vectors (Wehmeier Gene 165:149-150 (1995)) makes the process of recursive sequence recombination especially attractive by the techniques described above. Techniques for selection of antibiotic producing organisms can be used as described above; additionally, in some embodiments screening for a particular desired polyketide activity or compound is preferable.

4.14.5 ISOPRENOIDS

Isoprenoids result from cyclization of farnesyl pyrophosphate by sesquiterpene synthases. The diversity of isoprenoids is generated not by the backbone, but by control of cyclization. Cloned examples of isoprenoid synthesis genes include trichodiene synthase from *Fusarium sprorotrichioides*, pentalene synthase from *Streptomyces*, aristolochene synthase from *Penicillium roquefortii*, and epi-aristolochene synthase from *N. tabacum* (Cane, D.E. (1995). Isoprenoid antibiotics, pages 633-655, in "Genetics and Biochemistry of Antibiotic Production" edited by Vining, L.C. & Stuttard, C., published by Butterworth-Heinemann). Recursive sequence recombination of sesquiterpene synthases will be of use both in allowing expression of these enzymes in heterologous hosts (such as plants and industrial microbial strains) and in alteration of enzymes to change the cyclized product made. A large number of isoprenoids are active as antiviral, antibacterial, antifungal, herbicidal, insecticidal or cytostatic agents. Antibacterial and antifungal isoprenoids could thus be preferably screened for using the indicator cell type system described above, with the producing cell competing with bacteria or fungi for nutrients. Antiviral isoprenoids could be screened for preferably by their ability to confer resistance to viral attack on the producing cell.

4.14.6 BIOACTIVE PEPTIDE DERIVATIVES

Examples of bioactive non-ribosomally synthesized peptides include the antibiotics cyclosporin, pepstatin, actinomycin, gramicidin, depsipeptides, vancomycin, etc. These peptide derivatives are synthesized by complex enzymes rather than ribosomes.

Again, increasing the yield of such non-ribosomally synthesized peptide antibiotics has thus far been done by genetic identification of biosynthetic "bottlenecks" and over expression of specific enzymes (See, for example, p. 133-135 in "Genetics and Biochemistry of Antibiotic Production" edited by Vining, L.C. & Stuttard, C., published by Butterworth-Heinemann). Recursive sequence recombination of the enzyme clusters can be used to improve the yields of existing bioactive non-ribosomally made peptides in both natural and heterologous hosts.

Like polyketide synthases, peptide synthases are modular and multifunctional enzymes catalyzing condensation reactions between activated building blocks (in this case amino acids) followed by modifications of those building blocks (see Kleinkauf, H. and von Dohren, H. Eur. J. Biochem. 236:335-351 (1996)). Thus, as for polyketide synthases, recursive sequence recombination can also be used to alter peptide synthases: modifying the specificity of the amino acid recognized by each binding site on the enzyme and altering the activity or substrate specificities of sites that modify these amino acids to produce novel compounds with antibiotic activity. Other peptide antibiotics are made ribosomally and then post-translationally modified. Examples of this type of antibiotics are lantibiotics (produced by gram positive bacteria such *Staphylococcus*, *Streptomyces*, *Bacillus*, and *Actinoplanes*) and microcins (produced by *Enterobacteriaceae*). Modifications of the original peptide include (in lantibiotics) dehydration of serine and threonine, condensation of dehydroamino acids with cysteine, or simple N- and C-terminal blocking (microcins). For ribosomally made antibiotics both the peptide-encoding sequence and the modifying enzymes may have their expression levels modified by recursive sequence recombination. Again, this will lead to both increased levels of antibiotic synthesis, and by modulation of the levels of the modifying enzymes (and the sequence of the ribosomally synthesized peptide itself) novel antibiotics.

Screening can be done as for other antibiotics as described above, including competition with a sensitive (or even initially insensitive) microbial species. Use of competing bacteria that have resistances to the antibiotic being produced will select strongly either for greatly elevated levels of that antibiotic (so that it swamps out the resistance mechanism) or for novel derivatives of that antibiotic that are not

neutralized by the resistance mechanism.

4.14.7 POLYMERS

Several examples of metabolic engineering to produce biopolymers have been reported, including the production of the biodegradable plastic polyhydroxybutarate (PHB), and the polysaccharide xanthan gum. For a review, see Cameron et al. *Applied Biochem. Biotech.* 38:105-140 (1993). Genes for these pathways have been cloned, making them excellent candidates for the recursive sequence recombination techniques described above. Expression of such evolved genes in a commercially viable host such as *E. coli* is an especially attractive application of this technology.

Examples of starting materials for recursive sequence recombination include but are not limited to genes from bacteria such as *Alcaligenes*, *Zoogloea*, *Rhizobium*, *Bacillus*, and *Azobacter*, which produce polyhydroxyalkanoates (PHAs) such as polyhydroxybutyrate (PHB) intracellularly as energy reserve materials in response to stress. Genes from *Alcaligenes eutrophus* that encode enzymes catalyzing the conversion of acetoacetyl CoA to PHB have been transferred both to *E. coli* and to the plant *Arabidopsis thaliana* (Poirier et al. *Science* 256:520-523 (1992)). Two of these genes (*phbB* and *phbC*, encoding acetoacetyl-CoA reductase and PHB synthase respectively) allow production of PHE in *Arabidopsis*. The plants producing the plastic are stunted, probably because of adverse interactions between the new metabolic pathway and the plants' original metabolism (i.e., depletion of substrate from the mevalonate pathway). Improved production of PHB in plants has been attempted by localization of the pathway enzymes to organelles such as plastids. Other strategies such as regulation of tissue specificity, expression timing and cellular localization have been suggested to solve the deleterious effects of PHB expression in plants. The recursive sequence recombination techniques of the invention can be used to modify such heterologous genes as well as specific cloned interacting pathways (e.g., mevalonate), and to optimize PHB synthesis in industrial microbial strains, for example to remove the requirement for stresses (such as nitrogen limitation) in growth conditions.

Additionally, other microbial polyesters are made by different bacteria in which additional monomers are incorporated into the polymer (Peoples et al. in Novel Biodegradable Microbial Polymers, EA Dawes, ed., pp191-202 (1990)). Recursive sequence recombination of these genes or pathways singly or in combination into a heterologous host will allow the production of a variety of polymers with differing properties, including variation of the monomer subunit ratios in the polymer.

Another polymer whose synthesis may be manipulated by recursive sequence recombination is cellulose. The genes for cellulose biosynthesis have been cloned from *Agrobacterium tumefaciens* (Matthysse, A.G. et. al. J. Bacteriol. 177:1069-1075 (1995)). Recursive sequence recombination of this biosynthetic pathway could be used either to increase synthesis of cellulose, or to produce mutants in which alternative sugars are incorporated into the polymer.

4.14.8 CAROTENOIDS

Carotenoids are a family of over 600 terpenoids produced in the general isoprenoid biosynthetic pathway by bacteria, fungi and plants (for a review, see Armstrong, J. Bact. 176:4795-4802 (1994)). These pigments protect organisms against photooxidative damage as well as functioning as anti-tumor agents, free radical-scavenging anti-oxidants, and enhancers of the immune response. Additionally, they are used commercially in pigmentation of cultured fish and shellfish. Examples of carotenoids include but are not limited to myxobacton, spheroidene, spheroidenone, lutein, astaxanthin, violaxanthin, 4-ketorulene, myxoxanthophyll, echinenone, lycopene, zeaxanthin and its mono- and di- glucosides, alpha-, beta-, gamma- and sigma-carotene, beta-cryptoxanthin monoglucoside and neoxanthin.

Carotenoid synthesis is catalyzed by relatively small numbers of clustered genes: 11 different genes within 12 kb of DNA from *Myxococcus xanthus* (Botella et al. Eur. J. Biochem. 233:238-248 (1995)) and 8 genes within 9 kb of DNA from *Rhodobacter sphaeroides* (Lang et. al. J. Bact. 177:2064-2073 (1995)). In some microorganisms, such as *Thermus thermophilus*, these genes are plasmid-borne (Tabata et al. FEBS Letts 341:251-255 (1994)). These features make carotenoid synthetic pathways

especially attractive candidates for recursive sequence recombination.

Transfer of some carotenoid genes into heterologous organisms results in expression. For example, genes from *Erwina uredovora* and *Haematococcus pluvialis* will function together in *E. coli* (Kajiwara et al. *Plant Mol. Biol.* 29:343-352 (1995)). *E. herbicola* genes will function in *R. sphaeroides* (Hunter et al. *J. Bact.* 176:3692-3697 (1994)). However, some other genes do not; for example, *R. capsulatus* genes do not direct carotenoid synthesis in *E. coli* (Marrs, *J. Bact.* 146:1003-1012 (1981)).

In an embodiment of the invention, the recursive sequence recombination techniques of the invention can be used to generate variants in the regulatory and/or structural elements of genes in the carotenoid synthesis pathway, allowing increased expression in heterologous hosts. Indeed, traditional techniques have been used to increase carotenoid production by increasing expression of a rate limiting enzyme in *Thermus thermophilus* (Hoshino et al. *Appl. Environ. Micro.* 59:3150-3153 (1993)). Furthermore, mutation of regulatory genes can cause constitutive expression of carotenoid synthesis in actinomycetes, where carotenoid photoinducibility is otherwise unstable and lost at a relatively high frequency in some species (Kato et al. *Mol. Gen. Genet.* 247:387-390 (1995)). These are both mutations that can be obtained by recursive sequence recombination.

The recursive sequence recombination techniques of the invention as described above can be used to evolve one or more carotenoid synthesis genes in a desired host without the need for analysis of regulatory mechanisms. Since carotenoids are colored, a colorimetric assay in microtiter plates, or even on growth media plates, can be used for screening for increased production.

In addition to increasing expression of carotenoids, carotenogenic biosynthetic pathways have the potential to produce a wide diversity of carotenoids, as the enzymes involved appear to be specific for the type of reaction they will catalyze, but not for the substrate that they modify. For example, two enzymes from the marine bacterium *Agrobacterium aurantiacum* (*CrtW* and *CrtZ*) synthesize six different ketocarotenoids from beta-carotene (Misawa et al. *J. Bact.* 177:6576-6584 (1995)). This relaxed substrate specificity means that a diversity of substrates can be

transformed into an even greater diversity of products. Introduction of foreign carotenoid genes into a cell can lead to novel and functional carotenoid-protein complexes, for example in photosynthetic complexes (Hunter et al. J.Bact. 176:3692-3697 (1994)). Thus, the deliberate recombination of enzymes through the recursive sequence recombination techniques of the invention is likely to generate novel compounds. Screening for such compounds can be accomplished, for example, by the cell competition/survival techniques discussed above and by a colorimetric assay for pigmented compounds.

Another method of identifying new compounds is to use standard analytical techniques such as mass spectroscopy, nuclear magnetic resonance, high performance liquid chromatography, etc. Recombinant microorganisms can be pooled and extracts or media supernatants assayed from these pools. Any positive pool can then be subdivided and the procedure repeated until the single positive is identified ("sib-selection").

4.14.9 INDIGO BIOSYNTHESIS

Many dyes, i.e. agents for imparting color, are specialty chemicals with significant markets. As an example, indigo is currently produced chemically. However, nine genes have been combined in *E. coli* to allow the synthesis of indigo from glucose via the tryptophan/indole pathway (Murdock et al. Bio/Technology 11:381-386 (1993)). A number of manipulations were performed to optimize indigo synthesis: cloning of nine genes, modification of the fermentation medium and directed changes in two operons to increase reaction rates and catalytic activities of several enzymes.

Nevertheless, bacterially produced indigo is not currently an economic proposition. The recursive sequence recombination techniques of the instant invention could be used to optimize indigo synthesizing enzyme expression levels and catalytic activities, leading to increased indigo production, thereby making the process commercially viable and reducing the environmental impact of indigo manufacture. Screening for increased indigo production can be done by colorimetric assays of cultures in microtiter plates.

4.14.10 AMINO ACIDS

Amino acids of particular commercial importance include but are not limited to phenylalanine, monosodium glutamate, glycine, lysine, threonine, tryptophan and methionine. Backman et al. (Ann. NY Acad. Sci. 589:16-24 (1990)) disclosed the enhanced production of phenylalanine in *E. coli* via a systematic and downstream strategy covering organism selection, optimization of biosynthetic capacity, and development of fermentation and recovery processes. As described in Simpson et al. (Biochem Soc Trans, 23:381-387 (1995)), current work in the field of amino acid production is focused on understanding the regulation of these pathways in great molecular detail.

The recursive sequence recombination techniques of the instant invention would obviate the need for this analysis to obtain bacterial strains with higher secreted amino acid yields. Amino acid production could be optimized for expression using recursive sequence recombination of the amino acid synthesis and secretion genes as well as enzymes at the regulatory phosphoenolpyruvate branchpoint, from such organisms as *Serratia marcescens*, *Bacillus*, and the *Corynebacterium* -*Brevibacterium* group. In some embodiments of the invention, screening for enhanced production is preferably done in microtiter wells, using chemical tests well known in the art that are specific for the desired amino acid. Screening/selection for amino acid synthesis can also be done by using auxotrophic reporter cells that are themselves unable to synthesize the amino acid in question. If these reporter cells also produce a compound that stimulates the growth of the amino acid producer (this could be a growth factor, or even a different amino acid), then library cells that produce more amino acid will in turn receive more growth stimulant and will therefore grow more rapidly.

4.14.11 VITAMIN C SYNTHESIS

L-Ascorbic acid (vitamin C) is a commercially important vitamin with a world production of over 35,000 tons in 1984. Most vitamin C is currently manufactured

chemically by the Reichstein process, although recently bacteria have been engineered that are able to transform glucose to 2,5-keto-gluconic acid, and that product to 2- keto-L-idonic acid, the precursor to L-ascorbic acid (Boudrant, Enzyme Microb. Technol. 12:322-329 (1990)).

The efficiencies of these enzymatic steps in bacteria are currently low. Using the recursive sequence recombination techniques of the instant invention, the genes can be genetically engineered to create one or more operons followed by expression optimization of such a hybrid L-ascorbic acid synthetic pathway to result in commercially viable microbial vitamin C biosynthesis. In some embodiments, screening for enhanced L-ascorbic acid production is preferably done in microtiter plates, using assays well known in the art.

4.15 TEST FOR RESISTANCE TO DRUGS

4.15.1 FIND DRUGS THAT INDUCE RESISTANCE SLOWLY

A similar strategy can be used to simulate viral acquisition of drug resistance. The object is to identify drugs for which resistance can be acquired only slowly, if at all. The viruses to be evolved are those that cause infections in humans for which at least modestly effective drugs are available. Substrates for recombination can come from induced mutants, natural variants of the same viral strain or different viruses. If the target of the drug is known (e.g., nucleotide analogs which inhibit the reverse transcriptase gene of HIV), focused libraries containing variants of the target gene can be produced. Recombination of a viral genome with a library of fragments is usually performed *in vitro*. However, in situations in which the library of fragments constitutes variants of viral genomes or fragments that can be encompassed in such genomes, recombination can also be performed *in vivo*, e.g., by transfecting cells with multiple substrate copies (see Section V). For screening, recombinant viral genomes are introduced into host cells susceptible to infection by the virus and the cells are exposed to a drug effective against the virus (initially at low concentration). The cells

can be spun to remove any noninfected virus. After a period of infection, progeny viruses can be collected from the culture medium, the progeny viruses being enriched for viruses that have acquired at least partial resistance to the drug. Alternatively, virally infected cells can be plated in a soft agar lawn and resistant viruses isolated from plaques. Plaque size provides some indication of the degree of viral resistance.

Progeny viruses surviving screening are subject to additional rounds of recombination and screening at increased stringency until a predetermined level of drug resistance has been acquired. The predetermined level of drug resistance may reflect the maximum dosage of a drug practical to administer to a patient without intolerable side effects. The analysis is particularly valuable for investigating acquisition of resistance to various combination of drugs, such as the growing list of approved anti-HIV drugs (e.g., AZT, ddI, ddC, d4T, TIBO 82150, nevirapine, 3TC, crizivian and ritonavir).

4.15.2 METHOD TO EVOLVE YEAST STRAINS

Fragments are cloned into a YAC vector, and the resulting YAC library is transformed into competent yeast cells. Transformants containing a YAC are identified by selecting for a positive selection marker present on the YAC. The cells are allowed to recover and are then pooled. Thereafter, the cells are induced to sporulate by transferring the cells from rich medium, to nitrogen and carbon limiting medium. In the course of sporulation, cells undergo meiosis. Spores are then induced to mate by return to rich media. Optionally, ascii are lysed to liberate spores, so that the spores can mate with other spores originating from other ascii. Mating results in recombination between YACs bearing different inserts, and between YACs and natural yeast chromosomes. The latter can be promoted by irradiating spores with ultra violet light. Recombination can give rise to new phenotypes either as a result of genes expressed by fragments on the YACs or as a result of recombination with host genes, or both.

After induction of recombination between YACs and natural yeast chromosomes, YACs are often eliminated by selecting against a negative selection marker on the YACs. For example, YACs containing the marker URA3 can be selected against by propagation on media containing 5-fluro orotic acid. Any exogenous or altered genetic material that remains is contained within natural yeast chromosomes. Optionally, further rounds of recombination between natural yeast chromosomes can be performed after elimination of YACs. Optionally, the same or different library of YACs can be transformed into the cells, and the above steps repeated. By recursively repeating this process, the diversity of the population is increased prior to screening.

After elimination of YACs, yeast are then screened or selected for a desired property. The property can be a new property conferred by transferred fragments, such as production of an antibiotic. The property can also be an improved property of the yeast such as improved capacity to express or secrete an exogenous protein, improved recombinogenicity, improved stability to temperature or solvents, or other property required of commercial or research strains of yeast.

Yeast strains surviving selection/screening are then subject to a further round of recombination. Recombination can be exclusively between the chromosomes of yeast surviving selection/screening. Alternatively, a library of fragments can be introduced into the yeast cells and recombined with endogenous yeast chromosomes as before. This library of fragments can be the same or different from the library used in the previous round of transformation. For example, the YACs could contain a library of genomic DNA isolated from a pool of the improved strains obtained in the earlier steps. YACs are eliminated as before, followed by additional rounds of recombination and/or transformation with further YAC libraries. Recombination is followed by another round of selection/screening, as above.

Further rounds of recombination/screening can be performed as needed until a yeast

strain has evolved to acquire the desired property.

An exemplary scheme for evolving yeast by introduction of a YAC library is yeast containing an endogenous diploid genome and a YAC library of fragments representing variants of a sequence. The library is transformed into the cells to yield 100-1000 colonies per µg DNA. Most transformed yeast cells now harbor a single YAC as well as endogenous chromosomes. Meiosis is induced by growth on nitrogen and carbon limiting medium. In the course of meiosis the YACs recombine with other chromosomes in the same cell. Haploid spores resulting from meiosis mate and regenerated diploid forms. The diploid forms now harbor recombinant chromosomes, parts of which come from endogenous chromosomes and parts from YACs.

Optionally, the YACs can now be cured from the cells by selecting against a negative selection marker present on the YACs. Irrespective whether YACs are selected against, cells are then screened or selected for a desired property. Cells surviving selection/screening are transformed with another YAC library to start another stochastic &/or non-stochastic mutagenesis cycle.

4.15.3 EVOLVE YACs FOR TRANSFER INTO RECIPIENT STRAIN

These methods are based in part on the fact that multiple YACs can be harbored in the same yeast cell, and YAC-YAC recombination is known to occur (Green & Olson, Science 250, 94-98 1990)). Inter-YAC recombination provides a format for which families of homologous genes harbored on fragments of >20 kb can be stochastic &/or non-stochastic mutagenized in vivo.

The starting population of DNA fragments show sequence similarity with each other but differ as a result of for example, induced, allelic or species diversity. Often DNA fragments are known or suspected to encode multiple genes that function in a common pathway.

The fragments are cloned into a Yac and transformed into yeast, typically with positive selection for transformants. The transformants are induced to sporulate, as a result of which chromosomes undergo meiosis. The cells are then mated. Most of the resulting diploid cells now carry two YACs each having a different insert. These are again induced to sporulate and mated. The resulting cells harbor YACs of recombined sequence. The cells can then be screened or selected for a desired property. Typically, such selection occurs in the yeast strain used for stochastic &/or non-stochastic mutagenesis. However, if fragments being stochastic &/or non-stochastic mutagenized are not expressed in yeast, YACs can be isolated and transferred to an appropriate cell type in which they are expressed for screening. Examples of such properties include the synthesis or degradation of a desired compound, increased secretion of a desired gene product, or other detectable phenotype.

Preferably, the YAC library is transformed into haploid a and haploid a cells. These cells are then induced to mate with each other, i.e., they are pooled and induced to mate by growth on rich medium. The diploid cells, each carrying two YACs, are then transferred to sporulation medium. During sporulation, the cells undergo meiosis, and homologous chromosomes recombine. In this case, the genes harbored in the YACs will recombine, diversifying their sequences. The resulting haploid ascospores are then liberated from the ascii by enzymatic degradation of the ascus wall or other available means and the pooled liberated haploid ascospores are induced to mate by transfer to rich medium. This process is repeated for several cycles to increase the diversity of the DNA cloned into the YACs. The resulting population of yeast cells, preferably in the haploid state, are either screened for improved properties, or the diversified DNA is delivered to another host cell or organism for screening.

Cells surviving selection/screening are subjected to successive cycles of pooling, sporulation, mating and selection/screening until the desired phenotype has been observed. Recombination can be achieved simply by transferring cells from rich

medium to carbon and nitrogen limited medium to induce sporulation, and then returning the spores to rich media to induce mating. Ascii can be lysed to stimulate mating of spores originating from different ascii.

After YACs have been evolved to encode a desired property they can be transferred to other cell types. Transfer can be by protoplast fusion, or retransformation with isolated DNA. For example, transfer of YACs from yeast to mammalian cells is discussed by Monaco & Larin, Trends in BiotechnoloSy 12, 280-286 (1994); Montoliu et al., Reprod. Fertil. Dev. 6, 577-84 (1994); Lamb et al., Curr. Opin. Genet. Dev. 5, 342-8 (1995). An exemplary scheme for stochastic &/or non-stochastic mutagenesis a YAC fragment library in yeast is shown herein. A library of YAC fragments representing genetic variants are transformed into yeast that have diploid endogenous chromosomes. The transformed yeast continue to have diploid endogenous chromosomes, plus a single YAC. The yeast are induced to undergo meiosis and sporulate. The spores contain haploid genomes and are selected for those which contain a YAC, using the YAC selective marker. The spores are induced to mate generating diploid cells. The diploid cells now contain two YACs bearing different inserts as well as diploid endogenous chromosomes. The cells are again induced to undergo meiosis and sporulate. during meiosis, recombination occurs between the YAC inserts, and recombinant YACs are segregated to ascocytes. Some ascocytes thus contain haploid endogenous chromosomes plus a YAC chromosome with a recombinant insert. The ascocytes mature to spores, which can mate again generating diploid cells. Some diploid cells now possess a diploid complement of endogenous chromosomes plus two recombinant YACs. These cells can then be taken through further cycles of meiosis, sporulation and mating. In each cycle, further recombination occurs between YAC inserts and further recombinant forms of inserts are generated. After one or several cycles of recombination has occurred, cells can be tested for acquisition of a desired property. Further cycles of recombination, followed by selection, can then be performed in similar fashion.

**4.15.4 IN VIVO STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS
OF GENES BY THE RECURSIVE MATING OF YEAST CELLS
HARBORING HOMOLOGOUS GENES IN IDENTICAL LOCI**

A goal of DNA stochastic &/or non-stochastic mutagenesis is to mimic and expand the combinatorial capabilities of sexual recombination. In vitro DNA stochastic &/or non-stochastic mutagenesis succeeds in this process. However, by changing the mechanism of recombination and altering the conditions under which recombination occurs, naturally in vitro recombination methods may jeopardize intrinsic information in a DNA sequence that renders it "evolvable." Stochastic &/or non-stochastic mutagenesis in vivo by employing the natural crossing over mechanisms that occur during meiosis may access inherent natural sequence information and provide a means of creating higher quality stochastic &/or non-stochastic mutagenized libraries. Described here is a method for the in vivo stochastic &/or non-stochastic mutagenesis of DNA that utilizes the natural mechanisms of meiotic recombination and provides an alternative method for DNA stochastic &/or non-stochastic mutagenesis.

The basic strategy is to clone genes to be stochastic &/or non-stochastic mutagenized into identical loci within the haploid genome of yeast. The haploid cells are then recursively induced to mate and to sporulate. The process subjects the cloned genes to recursive recombination during recursive cycles of meiosis. The resulting stochastic &/or non-stochastic mutagenized genes are then screened in in situ or isolated and screened under different conditions.

For example, if one wished to reassemble a family of five lipase genes, the following provides a means of doing so in vivo.

The open reading frame of each lipase is amplified by the PCR such that each ORF is flanked by identical 3' and 5' sequences. The 5' flanking sequence is identical to a region within the 5' coding sequence of the *S. cerevisiae ura3* gene and the 3' flanking

sequence is identical to a region within the 3' of the *ura 3* gene. The flanking sequences are chosen such that homologous recombination of the PCR product with the *ura 3* gene results in the incorporation of the lipase gene and the disruption of the *ura 3* ORF. Both *S. cerevisiae* a and *haploid* cells are then transformed with each of the PCR amplified lipase ORFs, and cells having incorporated a lipase gene into the *ura 3* locus are selected by growth on 5 fluoro orotic acid (5FOA is lethal to cells expressing functional URA3). The result is 10 cell types, two different mating types each harboring one of the five lipase genes in the disrupted *ura 3* locus. These cells are then pooled and grown under conditions where mating between the a and *cells* are favored, e.g. in rich medium. Mating results in a combinatorial mixture of diploid cells having all 32 possible combinations of lipase genes in the two *ura 3* loci. The cells are then induced to sporulate by growth under carbon and nitrogen limited conditions. During sporulation the diploid cells undergo meiosis to form four (two a and two *)* haploid ascospores housed in an ascus.

During meiosis II of the sporulation process sister chromatids align and crossover. The lipase genes cloned into the *ura 3* loci will also align and recombine. Thus the resulting haploid ascospores will represent a library of cells each harboring a different possible chimeric lipase gene, each a unique result of the meiotic recombination of the two lipase genes in the original diploid cell. The walls of asci are degraded by treatment with zymolase to liberate and allow the mixing of the individual ascospores. This mixture is then grown under conditions that promote the mating of the a and *haploid* cells. It is important to liberate the individual ascospores, since mating will otherwise occur between the ascospores within an ascus.

Mixing of the haploid cells allows recombination between more than two lipase genes, enabling "poolwise recombination." Mating brings together new combinations of chimeric genes that can then undergo recombination upon sporulation. The cells are recursively cycled through sporulation, ascospore mixing, and mating until sufficient diversity has been generated by the recursive pairwise recombination of the

five lipase genes. The individual chimeric lipase genes either can be screened directly in the haploid yeast cells or transferred to an appropriate expression host.

The process is described above for lipases and yeast; however, any sexual organisms into which genes can be directed can be employed, and any genes, of course, could be substituted for lipases. This process is analogous to the method of stochastic &/or non-stochastic mutagenesis whole genomes by recursive pairwise mating. The diversity, however, in the whole genome case is distributed throughout the host genome rather than localized to specific loci.

4.15.5 USING YACs TO CLONE UNLINKED GENES BUT FUNCTIONALLY IMPORTANT GENES FROM ONE SPECIES INTO ANOTHER

Stochastic &/or non-stochastic mutagenesis of YACs is particularly amenable to transfer of unlinked but functionally related genes from one species to another, particularly where such genes have not been identified. Such is the case for several commercially important natural products, such as taxol. Transfer of the genes in the metabolic pathway to a different organism is often desirable because organisms naturally producing such compounds are not well suited for mass culturing.

Clusters of such genes can be isolated by cloning a total genomic library of DNA from an organisms producing a useful compound into a YAC library. The YAC library is then transformed into yeast. The yeast is sporulated and mated such that recombination occurs between YACs and/or between YACs and natural yeast chromosomes.

Selection/screening is then performed for expression of the desired collection of genes. If the genes encode a biosynthetic pathway, expression can be detected from the appearance of product of the pathway. Production of individual enzymes in the

pathway, or intermediates of the final expression product or capacity of cells to metabolize such intermediates indicates partial acquisition of the synthetic pathway. The original library or a different library can be introduced into cells surviving/selection screening, and further rounds of recombination and selection/screening can be performed until the end product of the desired metabolic pathway is produced.

4.15.6 YAC-YAC STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS

If a phenotype of interest can be isolated to a single stretch of genomic DNA less than 2 megabases in length, it can be cloned into a YAC and replicated in *S. cerevisiae*. The cloning of similar stretches of DNA from related hosts into an identical YAC results in a population of yeast cells each harboring a YAC having a homologous insert effecting a desired phenotype. The recursive breeding of these yeast cells allows the homologous regions of these YACs to recombine during meiosis, allowing genes, pathways, and clusters to recombine during each cycle of meiosis. After several cycles of mating and segregation, the YAC inserts are well stochastic &/or non-stochastic mutagenized. The now very diverse yeast library could then be screened for phenotypic improvements resulting from the stochastic &/or non-stochastic mutagenesis of the YAC inserts.

4.15.7 YAC-CHROMOSOME STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS

“Mitotic” recombination occurs during cell division and results from the recombination of genes during replication. This type of recombination is not limited to that between sister chromatids and can be enhanced by agents that induce recombination machinery, such as nicking chemicals and ultraviolet irradiation. Since it is often difficult to directly mate across a species barrier, it is possible to induce the recombination of homologous genes originating from different species by providing the target genes to a desired host organism as a YAC library. The genes harbored in

this library are then induced to recombine with homologous genes on the host chromosome by enhanced mitotic recombination. This process is carried out recursively to generate a library of diverse organisms and then screened for those having the desired phenotypic improvements. The improved subpopulation is then mated recursively as above to identify new strains having accumulated multiple useful genetic alterations.

4.15.8 ACCUMULATION OF MULTIPLE YACs HARBORING USEFUL GENES

The accumulation of multiple unlinked genes that are required for the acquisition or improvement of a given phenotype can be accomplished by the stochastic &/or non-stochastic mutagenesis of YAC libraries. Genomic DNA from organisms having desired phenotypes, such as ethanol tolerance, thermotolerance, and the ability to ferment pentose sugars are pooled, fragmented and cloned into several different YAC vectors, each having a different selective marker (his, ura, ade, etc). *S. cerevisiae* are transformed with these libraries, and selected for their presence (using selective media i.e uracil dropout media for the YAC containing the *ura3* selective marker) and then screened for having acquired or improved a desired phenotype.

Surviving cells are pooled, mated recursively, and selected for the accumulation of multiple YACs (by propagation in medium with multiple nutritional dropouts). Cells that acquire multiple YACs harboring useful genomic inserts are identified by further screening. Optimized strains can be used directly, however, due to the burden a YAC may pose to a cell, the relevant YAC inserts can be minimized, subcloned, and recombined into the host chromosome, to generate a more stable production strain.

4.15.9 CHOICE OF HOST SSF ORGANISM

One example use for the present invention is to create an improved yeast for the production of ethanol from lignocellulosic biomass. Specifically, a yeast strain with

improved ethanol tolerance and thermostability/thermotolerance is desirable. Parent yeast strains known for good behavior in a Simultaneous Saccharification and Fermentation (SSF) process are identified. These strains are combined with others known to possess ethanol," tolerance and/or thermostability.

S. cerevisiae is highly amenable to development for optimized SSF processes. It inherently possesses several traits for this use, including the ability to import and ferment a variety of sugars such as sucrose, glucose, galactose, maltose and maltriose. Also, yeast has the capability to flocculate, enabling recovery of the yeast biomass at the end of a fermentation cycle, and allowing its re-use in subsequent bioprocesses. This is an important property in that it optimizes the use of nutrients in the growth medium. *S. cerevisiae* is also highly amenable to laboratory manipulation, has highly characterized genetics and possesses a sexual reproductive cycle. *S. cerevisiae* may be grown under either aerobic or anaerobic conditions, in contrast to some other potential SSF organisms that are strict anaerobes (e.g. *Clostridium* spp.), making them very difficult to handle in the laboratory. *S. cerevisiae* are also "generally regarded as safe" ("GRAS"), and, due to its widespread use for the production of important comestibles for the general public (e.g. beer, wine, bread, etc), is generally familiar and well known. *S. cerevisiae* is commonly used in fermentative processes, and the familiarity in its handling by fermentation experts eases the introduction of novel improved yeast strains into the industrial setting. *S. cerevisiae* strains that previously have been identified as particularly good SSF organisms, for example, *S. cerevisiae* D₅A (ATCC200062) (South CR and Lynd LR. (1994) Appl. Biochem. Biotechnol. 45/46: 467-481; Ranatunga TD et al. (1997) Biotechnol. Lett. 19:1125-1127) can be used for starting materials. In addition, other industrially used *S. cerevisiae* strains are optionally used as host strains, particularly those showing desirable fermentative characteristics, such as *S. cerevisiae* Y567 (ATCC24858) (Sitton OC et al. (1979) Process Biochem. 14(9): 7-10; Sitton OC et al. (1981) Adv. Biotechnol. 2: 231-237; McMurrough I et al. (1971) Folia Microbiol. 16: 346-349) and *S. cerevisiae* ACA 174 (ATCC 60868) (Benitez T et al. (1983) Appl. Environ. Microbiol. 45: 1429-1436; Chem. Eng. J. 50: B 1 7-B22, 1992), which have been shown to have desirable traits for large-scale fermentation.

4.15.10 CHOICE OF ETHANOL TOLERANT STRAINS

Many strains of *S. cerevisiae* have been isolated from high-ethanol environments, and have survived in the ethanol-rich environment by adaptive evolution. For example, strains from Sherry wine aging ("Flor" strains) have evolved highly functional mitochondria to enable their survival in a high-ethanol environment. It has been shown that transfer of these wine yeast mitochondria to other strains increases the recipient's resistance to high ethanol concentration, as well as thermotolerance (Jimenez, J. and Benitez, T (1988) Curr. Genet. 13: 461-469). There are several flor strains deposited in the ATCC, for example *S. cerevisiae* MY91 (ATCC 201301), MY138 (ATCC 201302), C5 (ATCC 201298), ET7 (ATCC 201299), LA6 (ATCC 201300), OSB21 (ATCC 201303), F23 (*S. globosus* ATCC 90920). Also, several flor strains of *S. uvarum* and *Torulaspora pretoriensis* have been deposited. Other ethanol-tolerant wine strains include *S. cerevisiae* ACA 174 (ATCC 60868), 15% ethanol, and *S. cerevisiae* A54 (ATCC 90921), isolated from wine containing 18% (v/v) ethanol, and NRCC 202036 (ATCC 46534), also a wine yeast. Other *S. cerevisiae* ethanologens that additionally exhibit enhanced ethanol tolerance include ATCC 24858, ATCC 24858, G 3706 (ATCC 42594), NRRL Y-265 (ATCC 60593), and ATCC 24845 - ATCC 24860. A strain of *S. pastorianus* (*S. carlsbergensis* ATCC 2345) has high ethanol-tolerance (13% v/v). *S. cerevisiae* Sa28 (ATCC 26603), from Jamaican cane juice sample, produces high levels of alcohol from molasses, is sugar tolerant, and produces ethanol from wood acid hydrolyzate. Several of the listed strains, as well as additional strains can be used as starting materials for breeding ethanol tolerance.

4.15.11 CHOICE OF TEMPERATURE TOLERANT STRAINS

A few temperature tolerant strains have been reported, including the highly flocculent strain *S. pastorianus* SA 23 (*S. carlsbergensis* ATCC 26602), which produces ethanol at elevated temperatures, and *S. cerevisiae* Kyokai 7 (*S. sake*, ATCC 26422), a sake yeast tolerant to brief heat and oxidative stress. Ballesteros et al ((1991) Appl.

Biochem. Biotechnol. 28/29: 307-315) examined 27 strains of yeast for their ability to grow and ferment glucose in the 32-45°C temperature range, including *Saccharomyces*, *Kluyveromyces* and *Candida* spp. Of these, the best thermotolerant clones were *Kluyveromyces marxianus* LG and *Kluyveromyces fragilis* 2671 (Ballesteros et al (1993) Appl. Biochem. Biotechnol. 39/40: 201-211). *S. cerevisiae-pretoriensis* FDHI was somewhat thermotolerant, however was poor in ethanol tolerance. Recursive recombination of this strain with others that display ethanol tolerance can be used to acquire the thermotolerant characteristics of the strain in progeny which also display ethanol tolerance. *Candida acidothermophilum* (Issatchenlaa orientalis, ATCC 203 8 1) is a good SSF strain that also exhibits improved performance in ethanol production from lignocellulosic biomass at higher SSF temperatures than *S. cerevisiae* D₅A (Kadam, KL, Schmidt, SL (1997) Appl. Microbial. Biotechnol. 48: 709-713). This strain can also be a genetic contributor to an improved SSF strain.

4.15.12 STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS OF STRAINS

In those instances where strains are highly related, a recursive mating strategy may be pursued. For example, a population of haploid *S. cerevisiae* (a and) are mutagenized and screened for improved EtOH or thermal tolerance. The improved haploid subpopulation are mixed together and mated as a pool and induced to sporulate. The resulting haploid spores are freed by degrading the ascii wall and mixed. The freed spores are then induced to mate and sporulate recursively. This process is repeated a sufficient number of times to generate all possible mutant combinations. The whole genome stochastic &/or non-stochastic mutagenized population (haploid) is then screened for further EtOH or thermal tolerance.

When strains are not sufficiently related for recursive mating, formats based on protoplast fusion may be employed. Recursive and poolwise protoplast fusion can be performed to generate chimeric populations of diverse parental strains. The resultant

pool of progeny is selected and screened to identify improved ethanol and thermal tolerant strains.

Alternatively, a YAC-based Whole Genome Stochastic &/or non-stochastic mutagenesis format can be used. In this format, YACs are used to shuttle large chromosomal fragments between strains. As detailed earlier, recombination occurs between YACs or between YACs, and the host chromosomes. Genomic DNA from organisms having desired phenotypes are pooled, fragmented and cloned into several different YAC vectors, each having a different selective marker (his, ura, ade, etc). *S. cerevisiae* are transformed with these libraries, and selected for their presence (using selective media, i.e. uracil dropout media for the YAC containing the Ura 3 selective marker) and then screened for having acquired or improved a desired phenotype. Surviving cells are pooled, mated recursively (as above), and selected for the accumulation of multiple YACs (by propagation in medium with multiple nutritional dropouts). Cells that acquire multiple YACs harboring useful genomic inserts are identified by further screening (see below).

4.15.13 SELECTION FOR IMPROVED STRAINS

Having produced large libraries of novel strains by mutagenesis and recombination, a first task is to isolate those strains that possess improvements in the desired phenotypes. Identification of the organism libraries is facilitated where the desired key traits are selectable phenotypes. For example, ethanol has different effects on the growth rate of a yeast population, viability, and fermentation rate. Inhibition of cell growth and viability increases with ethanol concentration, but high fermentative capacity is only inhibited at higher ethanol concentrations. Hence, selection of growing cells in ethanol is a viable approach to isolate ethanol-tolerant strains. Subsequently, the selected strains may be analyzed for their fermentative capacity to produce ethanol. Provided that growth and media conditions are the same for all strains (parents and progeny), a hierarchy of ethanol tolerance may be constructed.

Simple selection schemes for identification of thermal tolerant and ethanol tolerant strains are available and, in this case, are based on those previously designed to identify potentially useful SSF strains. Selection of ethanol tolerance is performed by exposing the population to ethanol, then plating the population and looking for growth. Colonies capable of growing after exposure to ethanol can be re-exposed to a higher concentration of ethanol and the cycle repeated until the most tolerant strains are selected. In order to discern strains possessing heritable ethanol tolerance from with temporarily acquired adaptations, these cycles may be punctuated with cycles of growth in the absence of selection (e.g. no ethanol).

Alternatively, the mixed population can be grown directly at increasing concentrations of ethanol, and the most tolerant strains enriched (Aguilera and Benitez, 1986, Arch Microbiol 4:337-44). For example this enrichment could be carried out in a chemostat or turbidostat. Similar selections can be developed for thermal tolerance, in which strains are identified by their ability to grow after a heat treatment, or directly for growth at elevated temperatures (Ballesteros et al., 1991, Applied Biochem and Biotech, 28:307-315). The best strains identified by these selections will be assayed more thoroughly in subsequent screens for ethanol, thermal tolerance or other properties of interest.

In one aspect, organisms having increased ethanol tolerance are selected for. A population of natural *S. cerevisiae* isolates are mutagenized. This population is then grown under fermentor conditions under low initial ethanol concentrations. Once the culture has reached saturation, the culture is diluted into fresh medium having a slightly higher ethanol content. This process of successive dilution into medium of incrementally increasing ethanol concentration is continued until a threshold of ethanol tolerance is reached. The surviving mutant population having the highest ethanol tolerance are then pooled and their genomes recombined by any method noted herein. Enrichment could also be achieved by a continuous culture in a chemostat or

turbidostat in which temperature or ethanol concentrations are progressively elevated. The resulting stochastic &/or non-stochastic mutagenized population are then exposed once again to the enrichment strategy but at a higher starting medium ethanol concentration. This strategy is optionally applied for the enrichment of thermotolerant cells and for the enrichment of cells having combined thermo- and ethanol tolerance.

4.15.14 SCREENING FOR IMPROVED STRAINS

Strains showing viability in initial selections are assayed more quantitatively for improvements in the desired properties before being restochastic &/or non-stochastic mutagenized with other strains.

Progeny resulting from mutagenesis of a strain, or those pre-selected for their ethanol tolerance and/or thermostability, can be plated on non-selective agar. Colonies can be picked robotically into microtiter dishes and grown. Cultures are replicated to fresh microtiter plates, and the replicates are incubated under the appropriate stress condition(s). The growth or metabolic activity of individual clones may be monitored and ranked. Indicators of viability can range from the size of growing colonies on solid media, density of growing cultures, or color change of a metabolic activity indicator added to liquid media. Strains that show the greatest viability are then mixed and stochastic &/or non-stochastic mutagenized, and the resulting progeny are rescreened under more stringent conditions

4.15.15 DEVELOPMENT OF AYEAST STRAIN CAPABLE OF CONVERTING CELLULOSE TO MONOMERIC SUGARS

Once a strain of yeast exhibiting thermotolerance and ethanol tolerance is developed, the degradation of cellulose to monomeric sugars is provided by the inclusion to the host strain of an efficient cellulase degradation pathway.

Additional desirable characteristic can be useful to enhance the production of ethanol by the host. For example, inclusion of heterologous enzymes and pathways that broaden the substrate sugar range may be performed. "Tuning" of the strain can be

accomplished by the addition of various other traits, or the restoration of certain endogenous traits that are desirable, but lost during the recombination procedures.

4.15.16 CONFERRING OF CELLULASE ACTIVITY

A vast number of cellulases and cellulase degradation systems have been characterized from fungi, bacteria and yeast (see reviews by Beguin, P and Aubert, J-P (1994) FEMS Microbial. Rev. 13: 25-58; Ohima, K. et al. (1997) Biotechnol. Genet. Eng. Rev. 14: 365414). An enzymatic pathway required for efficient saccharification of cellulose involves the synergistic action of endoglucanases (endo-1,4--D-glucanases, EC 3. 2.1.4), exocellobiohydrolases (exo-1,4--D-glucanases, EC 3.2.1.91), and - glucosidases (cellobiases, 1,4--D-glucanases EC 3.2.1.21). The heterologous production of cellulase enzymes in the ethanologen would enable the saccharification of cellulose, producing monomeric sugars that may be used by the organism for ethanol production. There are several advantages to the heterologous expression of a functional cellulase pathway in the ethanologen. For example, the SSF process would eliminate the need for a separate bioprocess step for saccharification, and would ameliorate end-product inhibition of cellulase enzymes by accumulated intermediate and product sugars.

Naturally occurring cellulase pathways are inserted into the ethanologen, or one may choose to use custom improved "hybrid" cellulase pathways, employing the coordinate action of cellulases derived from different natural sources, including thermophiles.

Several cellulases from non-Saccharomyces have been produced and secreted from this organism successfully, including bacterial, fungal, and yeast enzymes, for example T. reesei CBH I ((Shoemaker (1994), in "The Cellulase System of Trichodenna reesei: Trichoderma strain improvement and Expression of Trichoderma cellulases in Yeast," Online, Pinner, UK, 593-600). It is possible to employ

straightforward metabolic engineering techniques to engender cellulase activity in *Saccharomyces*. Also, yeast have been forced to acquire elements of cellulose degradation pathways by protoplast fusion (e.g. intergeneric hybrids of *Saccharomyces cerevisiae* and *Zygosaccharomyces fermentati*, a cellobiase-producing yeast, have been created (Pina A, et. al. (1986) *Appl. Environ. Microbial.* 51: 995-1003). In general, any cellulase component enzyme that derives from a closely related yeast organism could be transferred by protoplast fusion. Cellobiases produced by a somewhat broader range of yeast may be accessed by whole genome stochastic &/or non-stochastic mutagenesis in one of its many formats (e.g. whole, fragmented, YAC-based).

Optimally, the cellulase enzymes to be used should exhibit good synergy, an appropriate level of expression and secretion from the host, good specific activity (i.e. resistance to host degradation factors and enzyme modification) and stability in the desired SSF environment. An example of a hybrid cellulose degradation pathway having excellent synergy includes the following enzymes: CBH I exocellobiohydrolase of *Trichoderma reesei*, the *Acidothermus cellulolyticus* E1 endoglucanase, and the *Thermomonospora fusca* E3 exocellulase (Baker, et. al. (1998) *Appl. Biochem. Biotechnol.* 70-72: 395-403).

It is suggested here that these enzymes (or improved mutants thereof) be considered for use in the SSF organism, along with a cellobiase (-glucosidase), such as that from *Candida peltata*. Other possible cellulase systems to be considered should possess particularly good activity against crystalline cellulose, such as the *T. reesei* cellulase system (Teeri, TT, et. al. (1998) *Biochem. Soc. Trans.* 26: 173-178), or possess particularly good thermostability characteristics (e.g. cellulase systems from thermophilic organisms, such as *Thermomonospora fusca* (Zhang, S., et. al. (1995) *Biochem.* 34: 3386-335).

A rational approach to the cloning of cellulases in the ethanologenic yeast host could be used. For example, known cellulase genes are cloned into expression cassettes utilizing *S. cerevisiae* promoter sequences, and the resultant linear fragments of DNA may be transformed into the recipient host by placing short yeast sequences at the termini to encourage site-specific integration into the genome. This is preferred to plasmidic transformation for reasons of genetic stability and maintenance of the transforming DNA.

If an entire cellulose degradative pathway were introduced, a selection could be implemented in an agar-plate-based format, and a large number of clones could be assayed for cellulase activity in a short period of time. For example, selection for an exocellulase may be accessible by providing a soluble oligocellulose substrate or carboxymethylcellulose (CMC) as a sole carbon source to the host, otherwise unable to grow on agar containing this sole carbon source. Clones producing active cellulase pathways would grow by virtue of their ability to produce glucose.

Alternatively, if the different cellulases were to be introduced sequentially, it would be useful to first introduce a cellobiase, enabling a selection using commercially available cellobiose as a sole carbon source. Several strains of *S. cerevisiae* that are able to grow on cellobiose have been created by introduction of a cellobiase gene (e.g. Rajoka MI, et. al. (1998) Floia Microbiol. (Praha) 43, 129-135; Skory, CD, et. al. (1996) Curr. Genet. 30, 417-422; D'Auria, S, et. al. (1996) Appl. Biochem. Biotechnol. 61, 157-166; Adam, AC, et. al. (1995) Yeast 11, 395-406; Adam, AC (1991) Cuff. Genet. 20, 5-8).

Subsequent transformation of this organism with CBHI exocellulase can be selected for by growth on a cellulose substrate such as carboxymethylcellulose (CMC). Finally, addition of an endoglucanase creates a yeast strain with improved crystalline degradation capacity.

4.15.17 CONFERRING OF PENTOSE SUGAR UTILIZATION

Inclusion of pentose sugar utilization pathways is an important facet to a potentially useful SSF organism. The successful expression of xylose sugar utilization pathways for ethanol production has been reported in *Saccharomyces* (e.g. Chen, ZD and Ho, NWY (1993) *Appl. Biochem. Biotechnol.* 39/40 135-147).

It would also be useful to accomplish L-arabinose substrate utilization for ethanol production in the *Saccharomyces* host. Yeast strains that utilize L-arabinose include some *Candida* and *Pichia* spp. (McMillan JD and Boynton BL (1994) *Appl. Biochem. Biotechnol.* 45-46: 569-584; Dien BS, et al. (1996) *Appl. Biochem. Biotechnol.* 57-58: 233- 242). Genes necessary for arabinose fermentation in *E. coli* could also be introduced by rational means (e.g. as has been performed previously in *Z. mobilis* (Deanda K, et. al. (1996) *Appl. Environ. Microbial.* 62: 4465-4470))

4.15.18 CONFERRING OF OTHER USEFUL ACTIVITES

Several other traits that are important for optimization of an SSF strain have been shown to be transferable to *S. cerevisiae*. Like thermal tolerance, cellulase activity and pentose sugar utilization, these traits may not normally be exhibited by *Saccharomyces* (or the particular strain of *Saccharomyces* being used as a host), and may be added by genetic means.

For example, expression of human muscle acylphosphatase in *S. cerevisiae* has been suggested to increase ethanol production (Rougei, G., et. al. (1996) *Biotechnol. App. Biochem.* 23: 273- 278).

It can occur that evolved stress-tolerant SSF strain acquire some undesirable

mutations in the course of the evolution strategy. Indeed, this is a pervasive problem in strain improvement strategies that rely on mutagenesis techniques, and can result in highly unstable or fragile production strains. It is possible to restore some of these desirable traits by rational methods such as cloning of specific genes that have been knocked out or negatively influenced in the previous rounds of strain improvement. The advantage to this approach is specificity- the offending gene may be targeted directly. The disadvantage is that it may be time-consuming and repetitious if several genes have been compromised, and it only addresses problems that have been characterized. A preferred (and more traditional) approach to the removal of undesirable/deleterious mutations is to back-cross the evolved strain to a desirable parent strain (e.g. the original "host" SSF strain). This strategy has been employed successfully throughout strain improvement where accessible (i.e. for organisms that have sexual cycles of reproduction). When lacking the advantage of a sexual process, it has been accomplished by using other methods, such as parasexual recombination or protoplast fusion.

For example, the ability to flocculate was conferred on a non- flocculating strain of *S. cerevisiae* by protoplast fusion with a flocculation competent *S. cerevisiae* (Watari, J., et. al (1990) Agric. Biol. Chem. 54: 1677-1681).

4.16 METHOD OF IN VIVO AND IN VITRO DNA SHUFFLING

4.16.1 Applications

Disclosed is a method of producing random polynucleotides by introducing two or more related polynucleotides into a suitable host cell such that a hybrid polynucleotide is generated by recombination and reductive reassortment. Also provided are vector and expression vehicles including such polynucleotides, polypeptides expressed by the hybrid polynucleotides and a method for screening for hybrid polypeptides

4.16.2 Experimental Applications

This invention relates generally to recombination and more specifically to a method for preparing polynucleotides encoding a polypeptide by a method of *in vivo* reassortment of polynucleotide sequences containing regions of partial homology, assembling the polynucleotides to form at least one polynucleotide and screening the polynucleotides for the production of polypeptide(s) having a useful property.

4.16.3 History

An exceedingly large number of possibilities exist for purposeful and random combinations of amino acids within a protein to produce useful hybrid proteins and their corresponding biological molecules encoding for these hybrid proteins, *i.e.*, DNA, RNA. Accordingly, there is a need to produce and screen a wide variety of such hybrid proteins for a useful utility, particularly widely varying random proteins.

The complexity of an active sequence of a biological macromolecule (*e.g.*, proteins, DNA) has been called its information content ("IC"), which has been defined as the resistance of the active protein to amino acid sequence variation (calculated from the minimum number of invariable amino acids (bits) required to describe a family of related sequences with the same function. Proteins that are more sensitive to random mutagenesis have a high information content.

Molecular biology developments, such as molecular libraries, have allowed the identification of quite a large number of variable bases, and even provide ways to select functional sequences from random libraries. In such libraries, most residues can be varied (although typically not all at the same time) depending on compensating changes in the context. Thus, while a 100 amino acid protein can contain only 2,000 different mutations, 20^{100} sequence combinations are possible.

Information density is the IC per unit length of a sequence. Active sites of enzymes tend to have a high information density. By contrast, flexible linkers of information in enzymes have a low information density.

Current methods in widespread use for creating alternative proteins in a library format are error-prone polymerase chain reactions and cassette mutagenesis, in which the specific region to be optimized is replaced with a synthetically mutagenized oligonucleotide. In both cases, a substantial number of mutant sites are generated around certain sites in the original sequence.

4.16.3.1 Error-prone PCR

Error-prone PCR uses low-fidelity polymerization conditions to introduce a low level of point mutations randomly over a long sequence. In a mixture of fragments of unknown sequence, error-prone PCR can be used to mutagenize the mixture. The published error-prone PCR protocols suffer from a low processivity of the polymerase. Therefore, the protocol is unable to result in the random mutagenesis of an average-sized gene. This inability limits the practical application of error-prone PCR. Some computer simulations have suggested that point mutagenesis alone may often be too gradual to allow the large-scale block changes that are required for continued and dramatic sequence evolution. Further, the published error-prone PCR protocols do not allow for amplification of DNA fragments greater than 0.5 to 1.0 kb, limiting their practical application. In addition, repeated cycles of error-prone PCR can lead to an accumulation of neutral mutations with undesired results, such as affecting a protein's immunogenicity but not its binding affinity

4.16.3.2 Oligonucleotide-Directed Mutagenesis

In oligonucleotide-directed mutagenesis, a short sequence is replaced with a synthetically mutagenized oligonucleotide. This approach does not generate combinations of distant mutations and is thus not combinatorial. The limited library size relative to the vast sequence length means that many rounds of selection are unavoidable for protein optimization. Mutagenesis with synthetic oligonucleotides requires sequencing of individual clones after each selection round followed by grouping them into families, arbitrarily choosing a single family, and reducing it to a consensus motif. Such motif is resynthesized and reinserted into a single gene followed by additional selection. This step process constitutes a statistical bottleneck,

is labor intensive, and is not practical for many rounds of mutagenesis.

Error-prone PCR and oligonucleotide-directed mutagenesis are thus useful for single cycles of sequence fine tuning, but rapidly become too limiting when they are applied for multiple cycles.

Another limitation of error-prone PCR is that the rate of down-mutations grows with the information content of the sequence. As the information content, library size, and mutagenesis rate increase, the balance of down-mutations to up-mutations will statistically prevent the selection of further improvements (statistical ceiling).

4.16.3.3 Cassette Mutagenesis

In cassette mutagenesis, a sequence block of a single template is typically replaced by a (partially) randomized sequence. Therefore, the maximum information content that can be obtained is statistically limited by the number of random sequences (*i.e.*, library size). This eliminates other sequence families which are not currently best, but which may have greater long term potential.

Also, mutagenesis with synthetic oligonucleotides requires sequencing of individual clones after each selection round. Thus, such an approach is tedious and impractical for many rounds of mutagenesis.

Thus, error-prone PCR and cassette mutagenesis are best suited, and have been widely used, for fine-tuning areas of comparatively low information content. One apparent exception is the selection of an RNA ligase ribozyme from a random library using many rounds of amplification by error-prone PCR and selection.

In nature, the evolution of most organisms occurs by natural selection and sexual reproduction. Sexual reproduction ensures mixing and combining of the genes in the offspring of the selected individuals. During meiosis, homologous chromosomes from the parents line up with one another and cross-over part way along their length, thus randomly swapping genetic material. Such swapping or shuffling of the DNA allows organisms to evolve more rapidly.

In recombination, because the inserted sequences were of proven utility in a homologous environment, the inserted sequences are likely to still have substantial information content once they are inserted into the new sequence.

4.16.3.4 Applied Molecular Evolution

The term Applied Molecular Evolution ("AME") means the application of an evolutionary design algorithm to a specific, useful goal. While many different library formats for AME have been reported for polynucleotides, peptides and proteins (phage, lacI and polysomes), none of these formats have provided for recombination by random cross-overs to deliberately create a combinatorial library.

Theoretically there are 2,000 different single mutants of a 100 amino acid protein. However, a protein of 100 amino acids has 20^{100} possible sequence combinations, a number which is too large to exhaustively explore by conventional methods. It would be advantageous to develop a system which would allow generation and screening of all of these possible combination mutations.

4.16.3.5 Reported *in vivo* Recombination Systems

Some workers in the art have utilized an *in vivo* site specific recombination system to generate hybrids of combine light chain antibody genes with heavy chain antibody genes for expression in a phage system. However, their system relies on specific sites of recombination and is limited accordingly. Simultaneous mutagenesis of antibody CDR regions in single chain antibodies (scFv) by overlapping extension and PCR have been reported.

Others have described a method for generating a large population of multiple hybrids using random *in vivo* recombination. This method requires the recombination of two different libraries of plasmids, each library having a different selectable marker. The method is limited to a finite number of recombinations equal to the number of selectable markers existing, and produces a concomitant linear increase in the number of marker genes linked to the selected sequence(s).

In vivo recombination between two homologous, but truncated, insect-toxin genes on a plasmid has been reported as a method of producing a hybrid gene. The *in vivo*

recombination of substantially mismatched DNA sequences in a host cell having defective mismatch repair enzymes, resulting in hybrid molecule formation has been reported.

4.16.4 Strategies

In one aspect this invention provides a method that utilizes the natural property of cells to recombine molecules and/or to mediate reductive processes that reduce the complexity of sequences and extent of repeated or consecutive sequences possessing regions of homology.

It is an object of the present invention to provide a method for generating hybrid polynucleotides encoding biologically active hybrid polypeptides with enhanced activities. In accomplishing these and other objects, there has been provided, in accordance with one aspect of the invention, a method for introducing polynucleotides into a suitable host cell and growing the host cell under conditions which produce a hybrid polynucleotide.

In another aspect of the invention, the invention provides a method for screening for biologically active hybrid polypeptides encoded by hybrid polynucleotides. The present method allows for the identification of biologically active hybrid polypeptides with enhanced biological activities.

Other objects, features and advantages of the present invention will become apparent from the following detailed description. It should be understood, however, that the detailed description and the specific examples, while indicating preferred embodiments of the invention, are given by way of illustration only, since various changes and modifications within the spirit and scope of the invention will become apparent to those skilled in the art from this detailed description.

4.16.5 Possible Uses

The invention described herein is directed to the use of repeated cycles of reductive reassortment, recombination and selection which allow for the directed molecular evolution of highly complex linear sequences, such as DNA, RNA or proteins

thorough recombination.

In vivo shuffling of molecules can be performed utilizing the natural property of cells to recombine multimers. While recombination *in vivo* has provided the major natural route to molecular diversity, genetic recombination remains a relatively complex process that involves 1) the recognition of homologies; 2) strand cleavage, strand invasion, and metabolic steps leading to the production of recombinant chiasma; and finally 3) the resolution of chiasma into discrete recombined molecules. The formation of the chiasma requires the recognition of homologous sequences.

4.16.5.1 Production of a Hybrid Polynucleotide

In a preferred embodiment, the invention relates to a method for producing a hybrid polynucleotide from at least a first polynucleotide and a second polynucleotide. The present invention can be used to produce a hybrid polynucleotide by introducing at least a first polynucleotide and a second polynucleotide which share at least one region of partial sequence homology into a suitable host cell. The regions of partial sequence homology promote processes which result in sequence reorganization producing a hybrid polynucleotide. The term "hybrid polynucleotide", as used herein, is any nucleotide sequence which results from the method of the present invention and contains sequence from at least two original polynucleotide sequences. Such hybrid polynucleotides can result from intermolecular recombination events which promote sequence integration between DNA molecules. In addition, such hybrid polynucleotides can result from intramolecular reductive reassortment processes which utilize repeated sequences to alter a nucleotide sequence within a DNA molecule.

The invention provides a means for generating hybrid polynucleotides which may encode biologically active hybrid polypeptides. In one aspect, the original polynucleotides encode biologically active polypeptides. The method of the invention produces new hybrid polypeptides by utilizing cellular processes which integrate the sequence of the original polynucleotides such that the resulting hybrid polynucleotide encodes a polypeptide demonstrating activities derived from the original biologically

active polypeptides. For example, the original polynucleotides may encode a particular enzyme from different microorganisms. An enzyme encoded by a first polynucleotide from one organism may, for example, function effectively under a particular environmental condition, *e.g.* high salinity. An enzyme encoded by a second polynucleotide from a different organism may function effectively under a different environmental condition, such as extremely high temperatures. A hybrid polynucleotide containing sequences from the first and second original polynucleotides may encode an enzyme which exhibits characteristics of both enzymes encoded by the original polynucleotides. Thus, the enzyme encoded by the hybrid polynucleotide may function effectively under environmental conditions shared by each of the enzymes encoded by the first and second polynucleotides, *e.g.*, high salinity and extreme temperatures.

4.16.5.1.1 Encoded Enzymes

Enzymes encoded by the original polynucleotides of the invention include, but are not limited to; oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases. A hybrid polypeptide resulting from the method of the invention may exhibit specialized enzyme activity not displayed in the original enzymes. For example, following recombination and/or reductive reassortment of polynucleotides encoding hydrolase activities, the resulting hybrid polypeptide encoded by a hybrid polynucleotide can be screened for specialized hydrolase activities obtained from each of the original enzymes, *i.e.* the type of bond on which the hydrolase acts and the temperature at which the hydrolase functions. Thus, for example, the hydrolase may be screened to ascertain those chemical functionalities which distinguish the hybrid hydrolase from the original hydrolyases, such as: (a) amide (peptide bonds), *i.e.* proteases; (b) ester bonds, *i.e.* esterases and lipases; (c) acetals, *i.e.*, glycosidases and, for example, the temperature, pH or salt concentration at which the hybrid polypeptide functions.

4.16.5.1.2 Sources Of The Original Polynucleotides

Sources of the original polynucleotides may be isolated from individual organisms ("isolates"), collections of organisms that have been grown in defined media ("enrichment cultures"), or, most preferably, uncultivated organisms ("environmental samples"). The use of a culture-independent approach to derive polynucleotides encoding novel bioactivities from environmental samples is most preferable since it allows one to access untapped resources of biodiversity.

"Environmental libraries" are generated from environmental samples and represent the collective genomes of naturally occurring organisms archived in cloning vectors that can be propagated in suitable prokaryotic hosts. Because the cloned DNA is initially extracted directly from environmental samples, the libraries are not limited to the small fraction of prokaryotes that can be grown in pure culture. Additionally, a normalization of the environmental DNA present in these samples could allow more equal representation of the DNA from all of the species present in the original sample. This can dramatically increase the efficiency of finding interesting genes from minor constituents of the sample which may be under-represented by several orders of magnitude compared to the dominant species.

For example, gene libraries generated from one or more uncultivated microorganisms are screened for an activity of interest. Potential pathways encoding bioactive molecules of interest are first captured in prokaryotic cells in the form of gene expression libraries. Polynucleotides encoding activities of interest are isolated from such libraries and introduced into a host cell. The host cell is grown under conditions which promote recombination and/or reductive reassortment creating potentially active biomolecules with novel or enhanced activities.

The microorganisms from which the polynucleotide may be prepared include prokaryotic microorganisms, such as Eubacteria and Archaebacteria, and lower eukaryotic microorganisms such as fungi, some algae and protozoa. Polynucleotides may be isolated from environmental samples in which case the nucleic acid may be recovered without culturing of an organism or recovered from one or more cultured organisms. In one aspect, such microorganisms may be extremophiles, such as hyperthermophiles, psychrophiles, psychrotrophs, halophiles, barophiles and acidophiles. Polynucleotides encoding enzymes isolated from extremophilic

microorganisms are particularly preferred. Such enzymes may function at temperatures above 100°C in terrestrial hot springs and deep sea thermal vents, at temperatures below 0°C in arctic waters, in the saturated salt environment of the Dead Sea, at pH values around 0 in coal deposits and geothermal sulfur-rich springs, or at pH values greater than 11 in sewage sludge. For example, several esterases and lipases cloned and expressed from extremophilic organisms show high activity throughout a wide range of temperatures and pHs.

4.16.5.1.3 Suitable Host Cells

Polynucleotides selected and isolated as hereinabove described are introduced into a suitable host cell. A suitable host cell is any cell which is capable of promoting recombination and/or reductive reassortment. The selected polynucleotides are preferably already in a vector which includes appropriate control sequences. The host cell can be a higher eukaryotic cell, such as a mammalian cell, or a lower eukaryotic cell, such as a yeast cell, or preferably, the host cell can be a prokaryotic cell, such as a bacterial cell. Introduction of the construct into the host cell can be effected by calcium phosphate transfection, DEAE-Dextran mediated transfection, or electroporation (Davis, L., Dibner, M., Battey, I., *Basic Methods in Molecular Biology*, (1986)).

As representative examples of appropriate hosts, there may be mentioned: bacterial cells, such as *E. coli*, *Streptomyces*, *Salmonella typhimurium*; fungal cells, such as yeast; insect cells such as *Drosophila S2* and *Spodoptera Sf9*; animal cells such as CHO, COS or Bowes melanoma; adenoviruses; and plant cells. The selection of an appropriate host is deemed to be within the scope of those skilled in the art from the teachings herein.

4.16.5.1.3.1 Mammalian Cell Culture Systems

With particular references to various mammalian cell culture systems that can be employed to express recombinant protein, examples of mammalian expression systems include the COS-7 lines of monkey kidney fibroblasts, described by

Gluzman, Cell, 23:175 (1981), and other cell lines capable of expressing a compatible vector, for example, the C127, 3T3, CHO, HeLa and BHK cell lines. Mammalian expression vectors will comprise an origin of replication, a suitable promoter and enhancer, and also any necessary ribosome binding sites, polyadenylation site, splice donor and acceptor sites, transcriptional termination sequences, and 5' flanking nontranscribed sequences. DNA sequences derived from the SV40 splice, and polyadenylation sites may be used to provide the required nontranscribed genetic elements.

Host cells containing the polynucleotides of interest can be cultured in conventional nutrient media modified as appropriate for activating promoters, selecting transformants or amplifying genes. The culture conditions, such as temperature, pH and the like, are those previously used with the host cell selected for expression, and will be apparent to the ordinarily skilled artisan. The clones which are identified as having the specified enzyme activity may then be sequenced to identify the polynucleotide sequence encoding an enzyme having the enhanced activity.

4.16.5.1.4 Generation Of Polynucleotides Encoding Biochemical Pathways

In another aspect, it is envisioned the method of the present invention can be used to generate novel polynucleotides encoding biochemical pathways from one or more operons or gene clusters or portions thereof. For example, bacteria and many eukaryotes have a coordinated mechanism for regulating genes whose products are involved in related processes. The genes are clustered, in structures referred to as "gene clusters," on a single chromosome and are transcribed together under the control of a single regulatory sequence, including a single promoter which initiates transcription of the entire cluster. Thus, a gene cluster is a group of adjacent genes that are either identical or related, usually as to their function. An example of a biochemical pathway encoded by gene clusters are polyketides. Polyketides are molecules which are an extremely rich source of bioactivities, including antibiotics (such as tetracyclines and erythromycin), anti-cancer agents (daunomycin), immunosuppressants (FK506 and rapamycin), and veterinary products (monensin). Many polyketides (produced by polyketide synthases) are valuable as therapeutic agents. Polyketide synthases are multifunctional enzymes that catalyze the

biosynthesis of an enormous variety of carbon chains differing in length and patterns of functionality and cyclization. Polyketide synthase genes fall into gene clusters and at least one type (designated type I) of polyketide synthases have large size genes and enzymes, complicating genetic manipulation and *in vitro* studies of these genes/proteins.

The ability to select and combine desired components from a library of polyketides, or fragments thereof, and postpolyketide biosynthesis genes for generation of novel polyketides for study is appealing. The method of the present invention makes it possible to facilitate the production of novel polyketide synthases through intermolecular recombination.

4.16.5.1.5 Gene Cluster DNA

Preferably, gene cluster DNA can be isolated from different organisms and ligated into vectors, particularly vectors containing expression regulatory sequences which can control and regulate the production of a detectable protein or protein-related array activity from the ligated gene clusters. Use of vectors which have an exceptionally large capacity for exogenous DNA introduction are particularly appropriate for use with such gene clusters and are described by way of example herein to include the f-factor (or fertility factor) of *E. coli*. This f-factor of *E. coli* is a plasmid which affect high-frequency transfer of itself during conjugation and is ideal to achieve and stably propagate large DNA fragments, such as gene clusters from mixed microbial samples. Once ligated into an appropriate vector, two or more vectors containing different polyketide synthase gene clusters can be introduced into a suitable host cell. Regions of partial sequence homology shared by the gene clusters will promote processes which result in sequence reorganization resulting in a hybrid gene cluster. The novel hybrid gene cluster can then be screened for enhanced activities not found in the original gene clusters.

Therefore, in a preferred embodiment, the present invention relates to a method for producing a biologically active hybrid polypeptide and screening such a

polypeptide for enhanced activity by:

introducing at least a first polynucleotide in operable linkage and a second polynucleotide in operable linkage, said at least first polynucleotide and second polynucleotide sharing at least one region of partial sequence homology, into a suitable host cell;

growing the host cell under conditions which promote sequence reorganization resulting in a hybrid polynucleotide in operable linkage;

expressing a hybrid polypeptide encoded by the hybrid polynucleotide;

screening the hybrid polypeptide under conditions which promote identification of enhanced biological activity; and

isolating the a polynucleotide encoding the hybrid polypeptide.

Methods for screening for various enzyme activities are known to those of skill in the art and discussed throughout the present specification. Such methods may be employed when isolating the polypeptides and polynucleotides of the present invention.

The term "isolated" means that material is removed from its original environment (e.g., the natural environment if it is naturally occurring). For example, a naturally-occurring polynucleotide or polypeptide present in a living animal is not isolated, but the same polynucleotide or polypeptide separated from some or all of the coexisting materials in the natural system, is isolated.

As used herein, the term "operably linked" refers to a linkage of polynucleotide elements in a functional relationship. A nucleic acid is "operably linked" when it is placed into a functional relationship with another nucleic acid sequence. For instance, a promoter or enhancer is operably linked to a coding sequence if it affects the transcription of the coding sequence. Operably linked means that the DNA sequences being linked are typically contiguous and, where necessary to join two protein coding regions, contiguous and in reading frame.

4.16.5.1.6 Expression Vectors

As representative examples of expression vectors which may be used there may be mentioned viral particles, baculovirus, phage, plasmids, phagemids, cosmids, fosmids, bacterial artificial chromosomes, viral DNA (*e.g.* vaccinia, adenovirus, foul pox virus, pseudorabies and derivatives of SV40), P1-based artificial chromosomes, yeast plasmids, yeast artificial chromosomes, and any other vectors specific for specific hosts of interest (such as bacillus, aspergillus and yeast). Thus, for example, the DNA may be included in any one of a variety of expression vectors for expressing a polypeptide. Such vectors include chromosomal, nonchromosomal and synthetic DNA sequences. Large numbers of suitable vectors are known to those of skill in the art, and are commercially available. The following vectors are provided by way of example; Bacterial: pQE vectors (Qiagen), pBluescript plasmids, pNH vectors, (lambda-ZAP vectors (Stratagene); ptrc99a, pKK223-3, pDR540, pRIT2T (Pharmacia); Eukaryotic: pXT1, pSG5 (Stratagene), pSVK3, pBPV, pMSG, pSVLSV40 (Pharmacia). However, any other plasmid or other vector may be used as long as they are replicable and viable in the host. Low copy number or high copy number vectors may be employed with the present invention.

A preferred type of vector for use in the present invention contains an f-factor origin replication. The f-factor (or fertility factor) in *E. coli* is a plasmid which effects high frequency transfer of itself during conjugation and less frequent transfer of the bacterial chromosome itself. A particularly preferred embodiment is to use cloning vectors, referred to as "fosmids" or bacterial artificial chromosome (BAC) vectors. These are derived from *E. coli* f-factor which is able to stably integrate large segments of genomic DNA. When integrated with DNA from a mixed uncultured environmental sample, this makes it possible to achieve large genomic fragments in the form of a stable "environmental DNA library."

Another preferred type of vector for use in the present invention is a cosmid vector. Cosmid vectors were originally designed to clone and propagate large segments of genomic DNA. Cloning into cosmid vectors is described in detail in Sambrook, *et al.*, Molecular Cloning A Laboratory Manual, Second Edition, Cold Spring Harbor Laboratory Press, 1989.

4.16.5.1.6.1 Expression Control Sequence

The DNA sequence in the expression vector is operatively linked to an appropriate expression control sequence(s) (promoter) to direct RNA synthesis. Particular named bacterial promoters include lacI, lacZ, T3, T7, gpt, lambda P_R, P_L and trp. Eukaryotic promoters include CMV immediate early, HSV thymidine kinase, early and late SV40, LTRs from retrovirus, and mouse metallothionein-I. Selection of the appropriate vector and promoter is well within the level of ordinary skill in the art. The expression vector also contains a ribosome binding site for translation initiation and a transcription terminator. The vector may also include appropriate sequences for amplifying expression. Promoter regions can be selected from any desired gene using CAT (chloramphenicol transferase) vectors or other vectors with selectable markers.

4.16.5.1.6.2 Selectable Marker Genes

In addition, the expression vectors preferably contain one or more selectable marker genes to provide a phenotypic trait for selection of transformed host cells such as dihydrofolate reductase or neomycin resistance for eukaryotic cell culture, or such as tetracycline or ampicillin resistance in *E. coli*.

Generally, recombinant expression vectors will include origins of replication and selectable markers permitting transformation of the host cell, *e.g.*, the ampicillin resistance gene of *E. coli* and *S. cerevisiae* TRP1 gene, and a promoter derived from a highly-expressed gene to direct transcription of a downstream structural sequence. Such promoters can be derived from operons encoding glycolytic enzymes such as 3-phosphoglycerate kinase (PGK), -factor, acid phosphatase, or heat shock proteins, among others. The heterologous structural sequence is assembled in appropriate phase with translation initiation and termination sequences, and preferably, a leader sequence capable of directing secretion of translated protein into the periplasmic space or extracellular medium.

The cloning strategy permits expression via both vector driven and endogenous promoters; vector promotion may be important with expression of genes whose

endogenous promoter will not function in *E. coli*.

4.16.5.1.7 Insertion Into a Vector or Plasmid

The DNA isolated or derived from microorganisms can preferably be inserted into a vector or a plasmid prior to probing for selected DNA. Such vectors or plasmids are preferably those containing expression regulatory sequences, including promoters, enhancers and the like. Such polynucleotides can be part of a vector and/or a composition and still be isolated, in that such vector or composition is not part of its natural environment. Particularly preferred phage or plasmid and methods for introduction and packaging into them are described in detail in the protocol set forth herein.

The selection of the cloning vector depends upon the approach taken, for example, the vector can be any cloning vector with an adequate capacity for multiply repeated copies of a sequence, or multiple sequences that can be successfully transformed and selected in a host cell. One example of such a vector is described in "Polycos vectors: a system for packaging filamentous phage and phagemid vectors using lambda phage packaging extracts", Alting-Mecs MA, Short JM, Gene, 1993 Dec 27, 137:1, 93-100. Propagation/maintenance can be by an antibiotic resistance carried by the cloning vector. After a period of growth, the naturally abbreviated molecules are recovered and identified by size fractionation on a gel or column, or amplified directly. The cloning vector utilized may contain a selectable gene that is disrupted by the insertion of the lengthy construct. As reductive reassortment progresses, the number of repeated units is reduced and the interrupted gene is again expressed and hence selection for the processed construct can be applied. The vector may be an expression/selection vector which will allow for the selection of an expressed product possessing desirable biologically properties. The insert may be positioned downstream of a functional promotor and the desirable property screened by appropriate means.

4.16.5.1.8 Reductive Reassortment

In vivo reassortment is focused on “inter-molecular” processes collectively referred to as “recombination” which in bacteria, is generally viewed as a “RecA-dependent” phenomenon. The present invention can rely on recombination processes of a host cell to recombine and re-assort sequences, or the cells ability to mediate reductive processes to decrease the complexity of quasi-repeated sequences in the cell by deletion. This process of “reductive reassortment” occurs by an “intra-molecular”, RecA-independent process.

Therefore, in another aspect of the present invention, novel polynucleotides can be generated by the process of reductive reassortment. The method involves the generation of constructs containing consecutive sequences (original encoding sequences), their insertion into an appropriate vector, and their subsequent introduction into an appropriate host cell. The reassortment of the individual molecular identities occurs by combinatorial processes between the consecutive sequences in the construct possessing regions of homology, or between quasi-repeated units. The reassortment process recombines and/or reduces the complexity and extent of the repeated sequences, and results in the production of novel molecular species. Various treatments may be applied to enhance the rate of reassortment. These could include treatment with ultra-violet light, or DNA damaging chemicals, and/or the use of host cell lines displaying enhanced levels of “genetic instability”. Thus the reassortment process may involve homologous recombination or the natural property of quasi-repeated sequences to direct their own evolution.

4.16.5.1.9 Repeated or “Quasi-Repeated” Sequences

Repeated or “quasi-repeated” sequences play a role in genetic instability. In the present invention, “quasi-repeats” are repeats that are not restricted to their original unit structure. Quasi-repeated units can be presented as an array of sequences in a construct; consecutive units of similar sequences. Once ligated, the junctions between the consecutive sequences become essentially invisible and the quasi-repetitive nature of the resulting construct is now continuous at the molecular level. The deletion process the cell performs to reduce the complexity of the resulting construct operates

between the quasi-repeated sequences. The quasi-repeated units provide a practically limitless repertoire of templates upon which slippage events can occur. The constructs containing the quasi-repeats thus effectively provide sufficient molecular elasticity that deletion (and potentially insertion) events can occur virtually anywhere within the quasi-repetitive units.

When the quasi-repeated sequences are all ligated in the same orientation, for instance head to tail or vice versa, the cell cannot distinguish individual units. Consequently, the reductive process can occur throughout the sequences. In contrast, when for example, the units are presented head to head, rather than head to tail, the inversion delineates the endpoints of the adjacent unit so that deletion formation will favor the loss of discrete units. Thus, it is preferable with the present method that the sequences are in the same orientation. Random orientation of quasi-repeated sequences will result in the loss of reassortment efficiency, while consistent orientation of the sequences will offer the highest efficiency. However, while having fewer of the contiguous sequences in the same orientation decreases the efficiency, it may still provide sufficient elasticity for the effective recovery of novel molecules. Constructs can be made with the quasi-repeated sequences in the same orientation to allow higher efficiency.

4.16.5.1.10 Assembly of Sequences in a Head to Tail Orientation

Sequences can be assembled in a head to tail orientation using any of a variety of methods, including the following:

- a) Primers that include a poly-A head and poly-T tail which when made single-stranded would provide orientation can be utilized. This is accomplished by having the first few bases of the primers made from RNA and hence easily removed RNaseH.
- b) Primers that include unique restriction cleavage sites can be utilized. Multiple sites, a battery of unique sequences, and repeated synthesis and ligation steps would be required.

- c) The inner few bases of the primer could be thiolated and an exonuclease used to produce properly tailed molecules.

4.16.5.1.11 The Recovery of the Re-assorted Sequences

The recovery of the re-assorted sequences relies on the identification of cloning vectors with a reduced RI. The re-assorted encoding sequences can then be recovered by amplification. The products are re-cloned and expressed. The recovery of cloning vectors with reduced RI can be effected by:

- 1) The use of vectors only stably maintained when the construct is reduced in complexity.
- 2) The physical recovery of shortened vectors by physical procedures. In this case, the cloning vector would be recovered using standard plasmid isolation procedures and size fractionated on either an agarose gel, or column with a low molecular weight cut off utilizing standard procedures.
- 3) The recovery of vectors containing interrupted genes which can be selected when insert size decreases.
- 4) The use of direct selection techniques with an expression vector and the appropriate selection.

Encoding sequences (for example, genes) from related organisms may demonstrate a high degree of homology and encode quite diverse protein products. These types of sequences are particularly useful in the present invention as quasi-repeats. However, while the examples illustrated below demonstrate the reassortment of nearly identical original encoding sequences (quasi-repeats), this process is not limited to such nearly identical repeats.

The following example demonstrates the method of the invention. Encoding nucleic acid sequences (quasi-repeats) derived from three (3) unique species are depicted. Each sequence encodes a protein with a distinct set of properties. Each of the sequences differs by a single or a few base pairs at a unique position in the

sequence which are designated "A", "B" and "C". The quasi-repeated sequences are separately or collectively amplified and ligated into random assemblies such that all possible permutations and combinations are available in the population of ligated molecules. The number of quasi-repeat units can be controlled by the assembly conditions. The average number of quasi-repeated units in a construct is defined as the repetitive index (RI).

Once formed, the constructs may, or may not be size fractionated on an agarose gel according to published protocols, inserted into a cloning vector, and transfected into an appropriate host cell. The cells are then propagated and "reductive reassortment" is effected. The rate of the reductive reassortment process may be stimulated by the introduction of DNA damage if desired. Whether the reduction in RI is mediated by deletion formation between repeated sequences by an "intra-molecular" mechanism, or mediated by recombination-like events through "inter-molecular" mechanisms is immaterial. The end result is a reassortment of the molecules into all possible combinations.

Optionally, the method comprises the additional step of screening the library members of the shuffled pool to identify individual shuffled library members having the ability to bind or otherwise interact (*e.g.*, such as catalytic antibodies) with a predetermined macromolecule, such as for example a proteinaceous receptor, peptide oligosaccharide, viron, or other predetermined compound or structure.

The displayed polypeptides, antibodies, peptidomimetic antibodies, and variable region sequences that are identified from such libraries can be used for therapeutic, diagnostic, research and related purposes (*e.g.*, catalysts, solutes for increasing osmolarity of an aqueous solution, and the like), and/or can be subjected to one or more additional cycles of shuffling and/or affinity selection. The method can be modified such that the step of selecting for a phenotypic characteristic can be other than of binding affinity for a predetermined molecule (*e.g.*, for catalytic activity, stability oxidation resistance, drug resistance, or detectable phenotype conferred upon a host cell).

4.16.5.1.12 Providing Antibodies Suitable for Affinity Interactions Screening

The present invention provides a method for generating libraries of displayed antibodies suitable for affinity interactions screening. The method comprises (1) obtaining first a plurality of selected library members comprising a displayed antibody and an associated polynucleotide encoding said displayed antibody, and obtaining said associated polynucleotide encoding for said displayed antibody and obtaining said associated polynucleotides or copies thereof, wherein said associated polynucleotides comprise a region of substantially identical variable region framework sequence, and (2) introducing said polynucleotides into a suitable host cell and growing the cells under conditions which promote recombination and reductive reassortment resulting in shuffled polynucleotides. CDR combinations comprised by the shuffled pool are not present in the first plurality of selected library members, said shuffled pool composing a library of displayed antibodies comprising CDR permutations and suitable for affinity interaction screening. Optionally, the shuffled pool is subjected to affinity screening to select shuffled library members which bind to a predetermined epitope (antigen) and thereby selecting a plurality of selected shuffled library members. Further, the plurality of selectively shuffled library members can be shuffled and screened iteratively, from 1 to about 1000 cycles or as desired until library members having a desired binding affinity are obtained.

4.16.5.1.13 Introduction of Mutations Into the Original Polynucleotides

In another aspect of the invention, it is envisioned that prior to or during recombination or reassortment, polynucleotides generated by the method of the present invention can be subjected to agents or processes which promote the introduction of mutations into the original polynucleotides. The introduction of such mutations would increase the diversity of resulting hybrid polynucleotides and polypeptides encoded therefrom. The agents or processes which promote mutagenesis can include, but are not limited to: (+)-CC-1065, or a synthetic analog such as (+)-CC-1065-(N3-Adenine), (see. Biochem. 31, 2822-2829 (1992)); a N-acetylated or deacetylated 4'-fluro-4-aminobiphenyl adduct capable of inhibiting DNA synthesis (see, for example, Carcinogenesis vol. 13, No. 5, 751-758 (1992); or a N-acetylated or deacetylated 4-aminobiphenyl adduct capable of inhibiting DNA

synthesis (see also, Id., 751-758); trivalent chromium, a trivalent chromium salt, a polycyclic aromatic hydrocarbon ("PAH") DNA adduct capable of inhibiting DNA replication, such as 7-bromomethyl-benz[*a*]anthracene ("BMA"), tris(2,3-dibromopropyl)phosphate ("Tris-BP"), 1,2-dibromo-3-chloropropane ("DBCP"), 2-bromoacrolein (2BA), benzo[*a*]pyrene-7,8-dihydrodiol-9-10-epoxide ("BPDE"), a platinum(II) halogen salt, N-hydroxy-2-amino-3-methylimidazo[4,5-*f*]-quinoline ("N-hydroxy-IQ"), and N-hydroxy-2-amino-1-methyl-6-phenylimidazo[4,5-*f*]-pyridine ("N-hydroxy-PhIP"). Especially preferred "means for slowing or halting PCR amplification consist of UV light (+)-CC-1065 and (+)-CC-1065-(N3-Adenine). Particularly encompassed means are DNA adducts or polynucleotides comprising the DNA adducts from the polynucleotides or polynucleotides pool, which can be released or removed by a process including heating the solution comprising the polynucleotides prior to further processing.

4.16.5.1.14 Production Of Hybrid Or Re-Assorted Polynucleotides

In another aspect the present invention is directed to a method of producing recombinant proteins having biological activity by treating a sample comprising double-stranded template polynucleotides encoding a wild-type protein under conditions according to the present invention which provide for the production of hybrid or re-assorted polynucleotides.

4.16.5.1.15 Shuffling a Population of Viral Genes of Viral Genomes

The invention also provides the use of polynucleotide shuffling to shuffle a population of viral genes (*e.g.*, capsid proteins, spike glycoproteins, polymerases, and proteases) or viral genomes (*e.g.*, paramyxoviridae, orthomyxoviridae, herpesviruses, retroviruses, reoviruses and rhinoviruses). In an embodiment, the invention provides a method for shuffling sequences encoding all or portions of immunogenic viral proteins to generate novel combinations of epitopes as well as novel epitopes created by recombination; such shuffled viral proteins may comprise epitopes or combinations of epitopes as well as novel epitopes created by recombination; such shuffled viral proteins may comprise epitopes or combinations of epitopes which are

likely to arise in the natural environment as a consequence of viral evolution; (*e.g.*, such as recombination of influenza virus strains).

4.16.5.1.16 Generation of Gene Therapy Vectors and Replication-Defective Gene Therapy Constructs

The invention also provides a method suitable for shuffling polynucleotide sequences for generating gene therapy vectors and replication-defective gene therapy constructs, such as may be used for human gene therapy, including but not limited to vaccination vectors for DNA-based vaccination, as well as anti-neoplastic gene therapy and other general therapy formats.

4.16.5.2 Definitions

The term "DNA shuffling" is used herein to indicate recombination between substantially homologous but non-identical sequences, in some embodiments DNA shuffling may involve crossover via non-homologous recombination, such as via cer/lox and/or flp/frt systems and the like.

The term "amplification" means that the number of copies of a polynucleotide is increased.

The term "identical" or "identity" means that two nucleic acid sequences have the same sequence or a complementary sequence. Thus, "areas of identity" means that regions or areas of a polynucleotide or the overall polynucleotide are identical or complementary to areas of another polynucleotide or the polynucleotide.

The term "corresponds to" is used herein to mean that a polynucleotide sequence is homologous (*i.e.*, is identical, not strictly evolutionarily related) to all or a portion of a reference polynucleotide sequence, or that a polypeptide sequence is identical to a reference polypeptide sequence. In contradistinction, the term "complementary to" is used herein to mean that the complementary sequence is homologous to all or a portion of a reference polynucleotide sequence. For illustration, the nucleotide sequence "TATAC" corresponds to a reference "TATAC" and is complementary to a reference sequence "GTATA."

Genetic instability, as used herein, refers to the natural tendency of highly repetitive

sequences to be lost through a process of reductive events generally involving sequence simplification through the loss of repeated sequences. Deletions tend to involve the loss of one copy of a repeat and everything between the repeats.

Quasi-repeated units, as used herein, refers to the repeats to be re-assorted and are by definition not identical. Indeed the method is proposed not only for practically identical encoding units produced by mutagenesis of the identical starting sequence, but also the reassortment of similar or related sequences which may diverge significantly in some regions. Nevertheless, if the sequences contain sufficient homologies to be reassorted by this approach, they can be referred to as "quasi-repeated" units.

Reductive reassortment, as used herein, refers to the increase in molecular diversity that is accrued through deletion (and/or insertion) events that are mediated by repeated sequences.

Repetitive Index (RI), as used herein, is the average number of copies of the quasi-repeated units contained in the cloning vector.

The term "related polynucleotides" means that regions or areas of the polynucleotides are identical and regions or areas of the polynucleotides are heterologous.

The term "population" as used herein means a collection of components such as polynucleotides, portions or polynucleotides or proteins. A "mixed population" means a collection of components which belong to the same family of nucleic acids or proteins (*i.e.*, are related) but which differ in their sequence (*i.e.*, are not identical) and hence in their biological activity.

The term "specific polynucleotide" means a polynucleotide having certain end points and having a certain nucleic acid sequence. Two polynucleotides wherein one polynucleotide has the identical sequence as a portion of the second polynucleotide but different ends comprises two different specific polynucleotides.

The following terms are used to describe the sequence relationships between two or more polynucleotides: "reference sequence," "comparison window," "sequence identity," "percentage of sequence identity," and "substantial identity." A "reference

"sequence" is a defined sequence used as a basis for a sequence comparison; a reference sequence may be a subset of a larger sequence, for example, as a segment of a full-length cDNA or gene sequence given in a sequence listing, or may comprise a complete cDNA or gene sequence. Generally, a reference sequence is at least 20 nucleotides in length, frequently at least 25 nucleotides in length, and often at least 50 nucleotides in length. Since two polynucleotides may each (1) comprise a sequence (*i.e.*, a portion of the complete polynucleotide sequence) that is similar between the two polynucleotides and (2) may further comprise a sequence that is divergent between the two polynucleotides, sequence comparisons between two (or more) polynucleotides are typically performed by comparing sequences of the two polynucleotides over a "comparison window" to identify and compare local regions of sequence similarity.

A "comparison window," as used herein, refers to a conceptual segment of at least 20 contiguous nucleotide positions wherein a polynucleotide sequence may be compared to a reference sequence of at least 20 contiguous nucleotides and wherein the portion of the polynucleotide sequence in the comparison window may comprise additions or deletions (*i.e.*, gaps) of 20 percent or less as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences. Optimal alignment of sequences for aligning a comparison window may be conducted by the local homology algorithm of Smith and Waterman (1981) Adv. Appl. Math. 2: 482 by the homology alignment algorithm of Needlemen and Wuncsch J. Mol. Biol. 48: 443 (1970), by the search of similarity method of Pearson and Lipman Proc. Natl. Acad. Sci. (U.S.A.) 85: 2444 (1988), by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package Release 7.0, Genetics Computer Group, 575 Science Dr., Madison, WI), or by inspection, and the best alignment (*i.e.*, resulting in the highest percentage of homology over the comparison window) generated by the various methods is selected.

The term "sequence identity" means that two polynucleotide sequences are identical (*i.e.*, on a nucleotide-by-nucleotide basis) over the window of comparison. The term "percentage of sequence identity" is calculated by comparing two optimally aligned

sequences over the window of comparison, determining the number of positions at which the identical nucleic acid base (*e.g.*, A, T, C, G, U, or I) occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison (*i.e.*, the window size), and multiplying the result by 100 to yield the percentage of sequence identity. This "substantial identity", as used herein, denotes a characteristic of a polynucleotide sequence, wherein the polynucleotide comprises a sequence having at least 80 percent sequence identity, preferably at least 85 percent identity, often 90 to 95 percent sequence identity, and most commonly at least 99 percent sequence identity as compared to a reference sequence of a comparison window of at least 25-50 nucleotides, wherein the percentage of sequence identity is calculated by comparing the reference sequence to the polynucleotide sequence which may include deletions or additions which total 20 percent or less of the reference sequence over the window of comparison.

"Conservative amino acid substitutions" refer to the interchangeability of residues having similar side chains. For example, a group of amino acids having aliphatic side chains is glycine, alanine, valine, leucine, and isoleucine; a group of amino acids having aliphatic-hydroxyl side chains is serine and threonine; a group of amino acids having amide-containing side chains is asparagine and glutamine; a group of amino acids having aromatic side chains is phenylalanine, tyrosine, and tryptophan; a group of amino acids having basic side chains is lysine, arginine, and histidine; and a group of amino acids having sulfur-containing side chains is cysteine and methionine. Preferred conservative amino acids substitution groups are : valine-leucine-isoleucine, phenylalanine-tyrosine, lysine-arginine, alanine-valine, and asparagine-glutamine.

The term "homologous" or "homeologous" means that one single-stranded nucleic acid nucleic acid sequence may hybridize to a complementary single-stranded nucleic acid sequence. The degree of hybridization may depend on a number of factors including the amount of identity between the sequences and the hybridization conditions such as temperature and salt concentrations as discussed later. Preferably the region of identity is greater than about 5 bp, more preferably the region of identity is greater than 10 bp.

The term "heterologous" means that one single-stranded nucleic acid sequence is unable to hybridize to another single-stranded nucleic acid sequence or its complement. Thus areas of heterology means that areas of polynucleotides or polynucleotides have areas or regions within their sequence which are unable to hybridize to another nucleic acid or polynucleotide. Such regions or areas are, for example areas of mutations.

The term "cognate" as used herein refers to a gene sequence that is evolutionarily and functionally related between species. For example but not limitation, in the human genome the human CD4 gene is the cognate gene to the mouse 3d4 gene, since the sequences and structures of these two genes indicate that they are highly homologous and both genes encode a protein which functions in signaling T cell activation through MHC class II-restricted antigen recognition.

The term "wild-type" means that the polynucleotide does not comprise any mutations. A "wild type" protein means that the protein will be active at a level of activity found in nature and will comprise the amino acid sequence found in nature.

The term "mutations" means changes in the sequence of a wild-type nucleic acid sequence or changes in the sequence of a peptide. Such mutations may be point mutations such as transitions or transversions. The mutations may be deletions, insertions or duplications.

In the polypeptide notation used herein, the left-hand direction is the amino terminal direction and the right-hand direction is the carboxy-terminal direction, in accordance with standard usage and convention. Similarly, unless specified otherwise, the left-hand end of single-stranded polynucleotide sequences is the 5' end; the left-hand direction of double-stranded polynucleotide sequences is referred to as the 5' direction. The direction of 5' to 3' addition of nascent RNA transcripts is referred to as the transcription direction; sequence regions on the DNA strand having the same sequence as the RNA and which are 5' to the 5' end of the RNA transcript are referred to as "upstream sequences"; sequence regions on the DNA strand having the same sequence as the RNA and which are 3' to the 3' end of the coding RNA transcript are referred to as "downstream sequences".

The term "naturally-occurring" as used herein as applied to the object refers to the fact that an object can be found in nature. For example, a polypeptide or polynucleotide sequence that is present in an organism (including viruses) that can be isolated from a source in nature and which has not been intentionally modified by man in the laboratory is naturally occurring. Generally, the term naturally occurring refers to an object as present in a non-pathological (un-diseased) individual, such as would be typical for the species.

The term "agent" is used herein to denote a chemical compound, a mixture of chemical compounds, an array of spatially localized compounds (*e.g.*, a VLSIPS peptide array, polynucleotide array, and/or combinatorial small molecule array), biological macromolecule, a bacteriophage peptide display library, a bacteriophage antibody (*e.g.*, scFv) display library, a polysome peptide display library, or an extract made form biological materials such as bacteria, plants, fungi, or animal (particular mammalian) cells or tissues. Agents are evaluated for potential activity as anti-neoplastics, anti-inflammatories or apoptosis modulators by inclusion in screening assays described hereinbelow. Agents are evaluated for potential activity as specific protein interaction inhibitors (*i.e.*, an agent which selectively inhibits a binding interaction between two predetermined polypeptides but which doe snot substantially interfere with cell viability) by inclusion in screening assays described hereinbelow.

As used herein, "substantially pure" means an object species is the predominant species present (*i.e.*, on a molar basis it is more abundant than any other individual macromolecular species in the composition), and preferably substantially purified fraction is a composition wherein the object species comprises at least about 50 percent (on a molar basis) of all macromolecular species present. Generally, a substantially pure composition will comprise more than about 80 to 90 percent of all macromolecular species present in the composition. Most preferably, the object species is purified to essential homogeneity (contaminant species cannot be detected in the composition by conventional detection methods) wherein the composition consists essentially of a single macromolecular species. Solvent species, small molecules (<500 Daltons), and elemental ion species are not considered macromolecular species.

As used herein the term "physiological conditions" refers to temperature, pH, ionic strength, viscosity, and like biochemical parameters which are compatible with a viable organism, and/or which typically exist intracellularly in a viable cultured yeast cell or mammalian cell. For example, the intracellular conditions in a yeast cell grown under typical laboratory culture conditions are physiological conditions.

Suitable *in vitro* reaction conditions for *in vitro* transcription cocktails are generally physiological conditions. In general, *in vitro* physiological conditions comprise 50-200 mM NaCl or KCl, pH 6.5-8.5, 20-45 C and 0.001-10 mM divalent cation (e.g., Mg⁺⁺, Ca⁺⁺); preferably about 150 mM NaCl or KCl, pH 7.2-7.6, 5 mM divalent cation, and often include 0.01-1.0 percent nonspecific protein (e.g., BSA). A non-ionic detergent (Tween, NP-40, Triton X-100) can often be present, usually at about 0.001 to 2%, typically 0.05-0.2% (v/v). Particular aqueous conditions may be selected by the practitioner according to conventional methods. For general guidance, the following buffered aqueous conditions may be applicable: 10-250 mM NaCl, 5-50 mM Tris HCl, pH 5-8, with optional addition of divalent cation(s) and/or metal chelators and/or non-ionic detergents and/or membrane fractions and/or anti-foam agents and/or scintillants.

"Specific hybridization" is defined herein as the formation of hybrids between a first polynucleotide and a second polynucleotide (e.g., a polynucleotide having a distinct but substantially identical sequence to the first polynucleotide), wherein substantially unrelated polynucleotide sequences do not form hybrids in the mixture.

As used herein, the term "single-chain antibody" refers to a polypeptide comprising a V_H domain and a V_L domain in polypeptide linkage, generally linked via a spacer peptide (e.g., [Gly-Gly-Gly-Gly-Ser]_x), and which may comprise additional amino acid sequences at the amino- and/or carboxy- termini. For example, a single-chain antibody may comprise a tether segment for linking to the encoding polynucleotide. As an example, a scFv is a single-chain antibody. Single-chain antibodies are generally proteins consisting of one or more polypeptide segments of at least 10 contiguous amino acids substantially encoded by genes of the immunoglobulin superfamily (e.g., see The Immunoglobulin Gene Superfamily, A.F. Williams and A.N. Barclay, in Immunoglobulin Genes, T. Honjo, F.W. Alt, and THE. Rabbits, eds., (1989) Academic

press: San Diego, CA, pp. 361-368, which is incorporated herein by reference), most frequently encoded by a rodent, non-human primate, avian, porcine bovine, ovine, goat, or human heavy chain or light chain gene sequence. A functional single-chain antibody generally contains a sufficient portion of an immunoglobulin superfamily gene product so as to retain the property of binding to a specific target molecule, typically a receptor or antigen (epitope).

As used herein, the term "complementarity-determining region" and "CDR" refer to the art-recognized term as exemplified by the Kabat and Chothia CDR definitions also generally known as supervariable regions or hypervariable loops (Chothia and Leks (1987) *J. Mol. Biol.* **196**; 901; Clothia *et al.* (1989) *Nature* **342**; 877; E.A. Kabat *et al.*, Sequences of Proteins of Immunological Interest (national Institutes of Health, Bethesda, MD) (1987); and Tramontano *et al.* (1990) *J. Mol. Biolog.* **215**; 175). Variable region domains typically comprise the amino-terminal approximately 105-115 amino acids of a naturally-occurring immunoglobulin chain (e.g., amino acids 1-110), although variable domains somewhat shorter or longer are also suitable for forming single-chain antibodies.

An immunoglobulin light or heavy chain variable region consists of a "framework" region interrupted by three hypervariable regions, also called CDR's. The extent of the framework region and CDR's have been precisely defined (see, "Sequences of Proteins of Immunological Interest," E. Kabat *et al.*, 4th Ed., U.S. Department of Health and human services, Bethesda, MD (1987)). The sequences of the framework regions of different light or heavy chains are relatively conserved within a specie. As used herein, a "human framework region" is a framework region that is substantially identical (about 85 or more, usually 90-95 or more) to the framework region of a naturally occurring human immunoglobulin. the framework region of an antibody, that is the combined framework regions of the constituent light and heavy chains, serves to position and align the CDR's. The CDR's are primarily responsible for binding to an epitope of an antigen.

As used herein, the term "variable segment" refers to a portion of a nascent peptide which comprises a random, pseudorandom, or defined kernal sequence. A variable segment" refers to a portion of a nascent peptide which comprises a random

pseudorandom, or defined kernal sequence. A variable segment can comprise both variant and invariant residue positions, and the degree of residue variation at a variant residue position may be limited: both options are selected at the discretion of the practitioner. Typically, variable segments are about 5 to 20 amino acid residues in length (e.g., 8 to 10), although variable segments may be longer and may comprise antibody portions or receptor proteins, such as an antibody fragment, a nucleic acid binding protein, a receptor protein, and the like.

As used herein, "random peptide sequence" refers to an amino acid sequence composed of two or more amino acid monomers and constructed by a stochastic or random process. A random peptide can include framework or scaffolding motifs, which may comprise invariant sequences.

As used herein "random peptide library" refers to a set of polynucleotide sequences that encodes a set of random peptides, and to the set of random peptides encoded by those polynucleotide sequences, as well as the fusion proteins contain those random peptides.

As used herein, the term "pseudorandom" refers to a set of sequences that have limited variability, such that, for example, the degree of residue variability at another position, but any pseudorandom position is allowed some degree of residue variation, however circumscribed.

As used herein, the term "defined sequence framework" refers to a set of defined sequences that are selected on a non-random basis, generally on the basis of experimental data or structural data; for example, a defined sequence framework may comprise a set of amino acid sequences that are predicted to form a β -sheet structure or may comprise a leucine zipper heptad repeat motif, a zinc-finger domain, among other variations. A "defined sequence kernal" is a set of sequences which encompass a limited scope of variability. Whereas (1) a completely random 10-mer sequence of the 20 conventional amino acids can be any of $(20)^{10}$ sequences, and (2) a pseudorandom 10-mer sequence of the 20 conventional amino acids can be any of $(20)^{10}$ sequences but will exhibit a bias for certain residues at certain positions and/or overall, (3) a defined sequence kernal is a subset of sequences if each residue position

was allowed to be any of the allowable 20 conventional amino acids (and/or allowable unconventional amino/imino acids). A defined sequence kernel generally comprises variant and invariant residue positions and/or comprises variant residue positions which can comprise a residue selected from a defined subset of amino acid residues), and the like, either segmentally or over the entire length of the individual selected library member sequence. Defined sequence kernels can refer to either amino acid sequences or polynucleotide sequences. Of illustration and not limitation, the sequences (NNK)₁₀ and (NNM)₁₀, wherein N represents A, T, G, or C; K represents G or T; and M represents A or C, are defined sequence kernels.

As used herein "epitope" refers to that portion of an antigen or other macromolecule capable of forming a binding interaction that interacts with the variable region binding body of an antibody. Typically, such binding interaction is manifested as an intermolecular contact with one or more amino acid residues of a CDR.

As used herein, "receptor" refers to a molecule that has an affinity for a given ligand. Receptors can be naturally occurring or synthetic molecules. Receptors can be employed in an unaltered state or as aggregates with other species. Receptors can be attached, covalently or non-covalently, to a binding member, either directly or via a specific binding substance. Examples of receptors include, but are not limited to, antibodies, including monoclonal antibodies and antisera reactive with specific antigenic determinants (such as on viruses, cells, or other materials), cell membrane receptors, complex carbohydrates and glycoproteins, enzymes, and hormone receptors.

As used herein "ligand" refers to a molecule, such as a random peptide or variable segment sequence, that is recognized by a particular receptor. As one of skill in the art will recognize, a molecule (or macromolecular complex) can be both a receptor and a ligand. In general, the binding partner having a smaller molecular weight is referred to as the ligand and the binding partner having a greater molecular weight is referred to as a receptor.

As used herein, "linker" or "spacer" refers to a molecule or group of molecules that connects two molecules, such as a DNA binding protein and a random peptide,

and serves to place the two molecules in a preferred configuration, e.g., so that the random peptide can bind to a receptor with minimal steric hindrance from the DNA binding protein.

4.16.5.3 Methodology

Nucleic acid shuffling is a method for *in vitro* or *in vivo* homologous recombination of pools of shorter or smaller polynucleotides to produce a polynucleotide or polynucleotides. Mixtures of related nucleic acid sequences or polynucleotides are subjected to sexual PCR to provide random polynucleotides, and reassembled to yield a library or mixed population of recombinant hybrid nucleic acid molecules or polynucleotides.

In contrast to cassette mutagenesis, only shuffling and error-prone PCR allow one to mutate a pool of sequences blindly (without sequence information other than primers).

4.16.5.3.1 Advantage of the Mutagenic Shuffling

The advantage of the mutagenic shuffling of this invention over error-prone PCR alone for repeated selection can best be explained with an example from antibody engineering.

4.16.5.3.2 Inverse Chain Reaction

This method differs from error-prone PCR, in that it is an inverse chain reaction. In error-prone PCR, the number of polymerase start sites and the number of molecules grows exponentially. However, the sequence of the polymerase start sites and the sequence of the molecules remains essentially the same. In contrast, in nucleic acid reassembly or shuffling of random polynucleotides the number of start sites and the number (but not size) of the random polynucleotides decreases over time. For polynucleotides derived from whole plasmids the theoretical endpoint is a single, large concatemeric molecule.

Since cross-overs occur at regions of homology, recombination will primarily occur

between members of the same sequence family. This discourages combinations of CDRs that are grossly incompatible (*e.g.*, directed against different epitopes of the same antigen). It is contemplated that multiple families of sequences can be shuffled in the same reaction. Further, shuffling generally conserves the relative order, such that, for example, CDR1 will not be found in the position of CDR2.

Rare shufflants will contain a large number of the best (*eg.* highest affinity) CDRs and these rare shufflants may be selected based on their superior affinity.

CDRs from a pool of 100 different selected antibody sequences can be permuted in up to 1006 different ways. This large number of permutations cannot be represented in a single library of DNA sequences. Accordingly, it is contemplated that multiple cycles of DNA shuffling and selection may be required depending on the length of the sequence and the sequence diversity desired.

Error-prone PCR, in contrast, keeps all the selected CDRs in the same relative sequence, generating a much smaller mutant cloud.

4.16.5.3.3 The Template Polynucleotide

The template polynucleotide which may be used in the methods of this invention may be DNA or RNA. It may be of various lengths depending on the size of the gene or shorter or smaller polynucleotide to be recombined or reassembled. Preferably, the template polynucleotide is from 50 bp to 50 kb. It is contemplated that entire vectors containing the nucleic acid encoding the protein of interest can be used in the methods of this invention, and in fact have been successfully used.

The template polynucleotide may be obtained by amplification using the PCR reaction (U.S. Patent No. 4,683,202 and 4,683,195) or other amplification or cloning methods. However, the removal of free primers from the PCR products before subjecting them to pooling of the PCR products and sexual PCR may provide more efficient results. Failure to adequately remove the primers from the original pool before sexual PCR can lead to a low frequency of crossover clones.

The template polynucleotide often should be double-stranded. A double-stranded

nucleic acid molecule is recommended to ensure that regions of the resulting single-stranded polynucleotides are complementary to each other and thus can hybridize to form a double-stranded molecule.

It is contemplated that single-stranded or double-stranded nucleic acid polynucleotides having regions of identity to the template polynucleotide and regions of heterology to the template polynucleotide may be added to the template polynucleotide, at this step. It is also contemplated that two different but related polynucleotide templates can be mixed at this step.

The double-stranded polynucleotide template and any added double-or single-stranded polynucleotides are subjected to sexual PCR which includes slowing or halting to provide a mixture of from about 5 bp to 5 kb or more. Preferably the size of the random polynucleotides is from about 10 bp to 1000 bp, more preferably the size of the polynucleotides is from about 20 bp to 500 bp.

4.16.5.3.4 Use of Double-Stranded Nucleic Acid Having Multiple Nicks

Alternatively, it is also contemplated that double-stranded nucleic acid having multiple nicks may be used in the methods of this invention. A nick is a break in one strand of the double-stranded nucleic acid. The distance between such nicks is preferably 5 bp to 5 kb, more preferably between 10 bp to 1000 bp. This can provide areas of self-priming to produce shorter or smaller polynucleotides to be included with the polynucleotides resulting from random primers, for example.

The concentration of any one specific polynucleotide will not be greater than 1% by weight of the total polynucleotides, more preferably the concentration of any one specific nucleic acid sequence will not be greater than 0.1% by weight of the total nucleic acid.

The number of different specific polynucleotides in the mixture will be at least about 100, preferably at least about 500, and more preferably at least about 1000.

4.16.5.3.5 Increasing the Heterogeneity of the Mixture of Polynucleotides

At this step single-stranded or double-stranded polynucleotides, either synthetic or natural, may be added to the random double-stranded shorter or smaller polynucleotides in order to increase the heterogeneity of the mixture of polynucleotides.

It is also contemplated that populations of double-stranded randomly broken polynucleotides may be mixed or combined at this step with the polynucleotides from the sexual PCR process and optionally subjected to one or more additional sexual PCR cycles.

Where insertion of mutations into the template polynucleotide is desired, single-stranded or double-stranded polynucleotides having a region of identity to the template polynucleotide and a region of heterology to the template polynucleotide may be added in a 20 fold excess by weight as compared to the total nucleic acid, more preferably the single-stranded polynucleotides may be added in a 10 fold excess by weight as compared to the total nucleic acid.

Where a mixture of different but related template polynucleotides is desired, populations of polynucleotides from each of the templates may be combined at a ratio of less than about 1:100, more preferably the ratio is less than about 1:40. For example, a backcross of the wild-type polynucleotide with a population of mutated polynucleotide may be desired to eliminate neutral mutations (e.g., mutations yielding an insubstantial alteration in the phenotypic property being selected for). In such an example, the ratio of randomly provided wild-type polynucleotides which may be added to the randomly provided sexual PCR cycle hybrid polynucleotides is approximately 1:1 to about 100:1, and more preferably from 1:1 to 40:1.

4.16.5.3.5.1 Denaturing and Re-annealing

The mixed population of random polynucleotides are denatured to form single-stranded polynucleotides and then re-annealed. Only those single-stranded polynucleotides having regions of homology with other single-stranded polynucleotides will re-anneal.

The random polynucleotides may be denatured by heating. One skilled in the art could determine the conditions necessary to completely denature the double-stranded nucleic acid. Preferably the temperature is from 80 °C to 100 °C, more preferably the temperature is from 90 °C to 96 °C. Other methods which may be used to denature the polynucleotides include pressure (36) and pH.

The polynucleotides may be re-annealed by cooling. Preferably the temperature is from 20 °C to 75 °C, more preferably the temperature is from 40 °C to 65 °C. If a high frequency of crossovers is needed based on an average of only 4 consecutive bases of homology, recombination can be forced by using a low annealing temperature, although the process becomes more difficult. The degree of renaturation which occurs will depend on the degree of homology between the population of single-stranded polynucleotides.

Renaturation can be accelerated by the addition of polyethylene glycol ("PEG") or salt. The salt concentration is preferably from 0 mM to 200 mM, more preferably the salt concentration is from 10 mM to 100 mM. The salt may be KCl or NaCl. The concentration of PEG is preferably from 0% to 20%, more preferably from 5% to 10%.

4.16.5.3.5.2 Incubation

The annealed polynucleotides are next incubated in the presence of a nucleic acid polymerase and dNTP's (*i.e.* dATP, dCTP, DGTP and dTTP). The nucleic acid polymerase may be the Klenow fragment, the Taq polymerase or any other DNA polymerase known in the art.

The approach to be used for the assembly depends on the minimum degree of homology that should still yield crossovers. If the areas of identity are large, Taq polymerase can be used with an annealing temperature of between 45-65 °C. If the areas of identity are small, Klenow polymerase can be used with an annealing temperature of between 20-30 °C. One skilled in the art could vary the temperature of annealing to increase the number of cross-overs achieved.

The polymerase may be added to the random polynucleotides prior to annealing, simultaneously with annealing or after annealing.

The cycle of denaturation, renaturation and incubation in the presence of polymerase is referred to herein as shuffling or reassembly of the nucleic acid. This cycle is repeated for a desired number of times. Preferably the cycle is repeated from 2 to 50 times, more preferably the sequence is repeated from 10 to 40 times.

4.16.5.3.6 The Resulting Nucleic Acid

The resulting nucleic acid is a larger double-stranded polynucleotide of from about 50 bp to about 100 kb, preferably the larger polynucleotide is from 500 bp to 50 kb.

This larger polynucleotides may contain a number of copies of a polynucleotide having the same size as the template polynucleotide in tandem. This concatemeric polynucleotide is then denatured into single copies of the template polynucleotide. The result will be a population of polynucleotides of approximately the same size as the template polynucleotide. The population will be a mixed population where single or double-stranded polynucleotides having an area of identity and an area of heterology have been added to the template polynucleotide prior to shuffling.

These polynucleotides are then cloned into the appropriate vector and the ligation mixture used to transform bacteria.

It is contemplated that the single polynucleotides may be obtained from the larger concatemeric polynucleotide by amplification of the single polynucleotide prior to cloning by a variety of methods including PCR (U.S. Patent No. 4,683,195 and 4,683,202), rather than by digestion of the concatemer.

4.16.5.3.7 Vectors Used for Cloning

The vector used for cloning is not critical provided that it will accept a polynucleotide of the desired size. If expression of the particular polynucleotide is desired, the cloning vehicle should further comprise transcription and translation signals next to the site of insertion of the polynucleotide to allow expression of the polynucleotide in

the host cell. Preferred vectors include the pUC series and the pBR series of plasmids.

4.16.5.3.8 The Resulting Bacterial Population

The resulting bacterial population will include a number of recombinant polynucleotides having random mutations. This mixed population may be tested to identify the desired recombinant polynucleotides. The method of selection will depend on the polynucleotide desired.

For example, if a polynucleotide which encodes a protein with increased binding efficiency to a ligand is desired, the proteins expressed by each of the portions of the polynucleotides in the population or library may be tested for their ability to bind to the ligand by methods known in the art (*i.e.* panning, affinity chromatography). If a polynucleotide which encodes for a protein with increased drug resistance is desired, the proteins expressed by each of the polynucleotides in the population or library may be tested for their ability to confer drug resistance to the host organism. One skilled in the art, given knowledge of the desired protein, could readily test the population to identify polynucleotides which confer the desired properties onto the protein.

It is contemplated that one skilled in the art could use a phage display system in which fragments of the protein are expressed as fusion proteins on the phage surface (Pharmacia, Milwaukee WI). The recombinant DNA molecules are cloned into the phage DNA at a site which results in the transcription of a fusion protein a portion of which is encoded by the recombinant DNA molecule. The phage containing the recombinant nucleic acid molecule undergoes replication and transcription in the cell. The leader sequence of the fusion protein directs the transport of the fusion protein to the tip of the phage particle. Thus the fusion protein which is partially encoded by the recombinant DNA molecule is displayed on the phage particle for detection and selection by the methods described above.

4.16.5.3.9 Cycles of Nucleic Acid Shuffling

It is further contemplated that a number of cycles of nucleic acid shuffling may be conducted with polynucleotides from a sub-population of the first population, which sub-population contains DNA encoding the desired recombinant protein. In this manner, proteins with even higher binding affinities or enzymatic activity could be achieved.

It is also contemplated that a number of cycles of nucleic acid shuffling may be conducted with a mixture of wild-type polynucleotides and a sub-population of nucleic acid from the first or subsequent rounds of nucleic acid shuffling in order to remove any silent mutations from the sub-population.

4.16.5.3.10 The Starting Nucleic Acid

Any source of nucleic acid, in purified form can be utilized as the starting nucleic acid. Thus the process may employ DNA or RNA including messenger RNA, which DNA or RNA may be single or double stranded. In addition, a DNA-RNA hybrid which contains one strand of each may be utilized. The nucleic acid sequence may be of various lengths depending on the size of the nucleic acid sequence to be mutated. Preferably the specific nucleic acid sequence is from 50 to 50000 base pairs. It is contemplated that entire vectors containing the nucleic acid encoding the protein of interest may be used in the methods of this invention.

The nucleic acid may be obtained from any source, for example, from plasmids such a pBR322, from cloned DNA or RNA or from natural DNA or RNA from any source including bacteria, yeast, viruses and higher organisms such as plants or animals. DNA or RNA may be extracted from blood or tissue material. The template polynucleotide may be obtained by amplification using the polynucleotide chain reaction (PCR) (U.S. Patent no. 4,683,202 and 4,683,195). Alternatively, the polynucleotide may be present in a vector present in a cell and sufficient nucleic acid may be obtained by culturing the cell and extracting the nucleic acid from the cell by methods known in the art.

Any specific nucleic acid sequence can be used to produce the population of hybrids by the present process. It is only necessary that a small population of hybrid sequences of the specific nucleic acid sequence exist or be created prior to the present